# SPRACHRAUMANALYSE MIT HILFE EINER PHONETISCHEN ONTOLOGIE

Robert Engsterhold
Marburg 2020

Dieses Buch stellt eine leicht überarbeitete Version der gleichnamigen Dissertationsschrift dar, die 2019 vom Fachbereich Germanistik und Kunstwissenschaften der Philipps-Universität Marburg angenommen wurde.

Angenommen am:
14.12.2018

Tag der Disputation:
29.1.2019

Betreuer/Erstgutachter:
Prof. Dr. Jürgen Erich Schmidt

Zweitgutachter:
Prof. Dr. Joachim Herrgen

## ABSTRACT

This thesis describes a quantitative analysis of a language area based on an ontology. It utilizes the digitized data of *The Linguistic Atlas of the Middle Rhine* (MRhSA), which was made available by the REDE project. An Ontology (*phonOntology*) was developed to further enrich the IPA-annotated observations of the atlas by breaking down each sound into its sound-properties. This is possible due the inference capabilities of the ontology, thus allowing for the creation of quantitative datasets that contain representative feature vectors for each place covered by the MRhSA. The study starts with a comprehensive dataset including all of the observations for an older generation of speakers interviewed during the field work for the MRhSA. In the next step, subsets of these are extracted on the basis of which datasets associated with the Middle High German and West Germanic reference systems are generated. These datasets are first analyzed via clustering algorithms and then evaluated based on different stability metrics. The results show a structural difference between the northern region called Moselle Franconian and the southern region called Rhine Franconian. This difference is not bound to specific isoglosses, but can be found within the sound properties themselves. It is possible to conduct an apparent-time analysis because the MRhSA also offers a second dataset that takes a second generation of younger speakers into account. The results of this analysis show that the structural difference is still in place; however, there are measurable normalization tendencies between these dialects, which may be attributable to a closer approximation to Standard German.



## ZUSAMMENFASSUNG

Diese Arbeit beschreibt eine quantitative Sprachraumanalyse auf Basis einer Ontologie. Ausgangspunkt sind dabei die im REDE-Projekt verfügbaren digitalisierten Kartendaten des Mittelrheinischen Sprachatlas (MRhSA). Damit die in IPA annotierten Lautobservationen genauer beschrieben werden können, wurde eine Ontologie (*phonOntologie*) entwickelt, welche die in IPA definierten Laute mit den zugehörigen phonetischen Eigenschaften in Beziehung setzt. Mittels Inferenz ist es möglich, automatisch aus den IPA-Lauten die entsprechenden, in der Ontologie definierten phonetischen Eigenschaften zu erhalten. Durch diese Technik und einer anschließenden Quantifizierung lassen sich Datensets konstruieren, die zu jedem Ort einen repräsentativen Lauteigenschaften-Vektor enthalten. Die Konstruktion der Datensets geschieht unter Berücksichtigung historischer Lautklassen. Die so generierten Datensets bilden ein Datenset zu allen Lauten, den Lauten zu den mittelhochdeutschen Langvokalen, mittelhochdeutschen Kurzvokalen und den westgermanischen Konsonanten. Diese Datensets werden anschließend mittels drei populärer Clusteralgorithmen geclustert und anhand von Stabilitätsmetriken bewertet. Es zeigt sich mittels dieser Analysen, dass es eine deutliche

strukturelle Trennung zwischen dem Moselfränkischen und dem Rheinfränkischen gibt. Zusätzlich tun sich bei höheren Clusterings noch Unterregionen auf, die auch auf strukturelle Phänomene hindeuten. Da im MRhSA neben dem Hauptdatenset, das auf den Aufnahmen von älteren Sprechern („ältere Generation“) basiert, auch ein zweites Datenset, mit Aufnahmen zu jüngeren Sprechern („jüngere Generation“) vorliegt, kann in einer „apparent-time“-Analyse, der Wandel zwischen den beiden Generationen beschrieben werden. Es zeigt sich, dass obwohl es deutliche Normalisierungstendenzen gibt, der strukturelle Unterschied gewahrt bleibt. Diese Normalisierung liegt wahrscheinlich an einer Annäherung an das Standarddeutsch in beiden Regionen.

# DANKSAGUNG

Diese Arbeit bietet die Sicht eines Informatikers auf ein Kernthema der Linguistik. Dafür, dass mir das ermöglicht wurde, möchte ich ganz besonders Herrn Professor Dr. Jürgen Erich Schmidt danken. Er hat mich, als ich neu am Forschungszentrum Deutscher Sprachatlas war, unter seine Fittiche genommen und mir geduldig (und wiederholt) die Grundlagen der Dialektologie erklärt, so dass sich das Thema und der Ansatz dieser Arbeit herausbilden konnte. Für diesen Ansatz und die Expertise, die ich in die Sprachraumforschung einbringen darf, möchte ich Herrn Professor Dr. Heinrich Herre danken, der mir während meines Studiums in Leipzig den faszinierenden Bereich der Ontologie nahebrachte; ein Teilgebiet der Informatik, das mich seitdem nicht mehr losgelassen hat. Weiterhin danke ich Herrn Professor Dr. Joachim Herrgen dafür, dass er mir die Geschichte des MRhSA nahebrachte und Herrn Professor Dr. Michael Cysouw für neue Gedankenansätze und Expertise in der quantitativen Linguistik.

Dafür, dass sie, während des Schreibens dieser Arbeit, die meisten Ablenkungen von mir ferngehalten haben, danke ich Herrn Professor Dr. Roland Kehrein und natürlich Dennis Bock, der diese Ablenkungen übernehmen musste.

Des Weiteren danke ich ganz besonders Hanna Fischer, Simon Kasper, Jeffrey Pheiff und Tillmann Pistor dafür, dass sie diese Arbeit Korrektur gelesen haben, mir immer bei linguistischen Fragestellungen helfen und für die erfrischende und erheiternde Dynamik, die sich zwischen uns in den letzten Jahren entwickelt hat. Außerdem danke ich noch Judith Hauff dafür, dass sie die Raumeinteilungen für die in dieser Arbeit verwendete Grundkarte in REDE gezeichnet hat.

# INHALTSVERZEICHNIS

# ABBILDUNGSVERZEICHNIS

## TABELLENVERZEICHNIS

# ABKÜRZUNGEN

| | |
|---|---|
| GMM | Gaussian Mixture Model |
| IPA | Internationales Phonetisches Alphabet |
| Mhd | Mittelhochdeutsch |
| MRhSA | Mittelrheinischer Sprachatlas |
| OWL | Web Ontology Language |
| RDF | Resource Description Framework |
| RDFS | Resource Description Framework Schema |
| W3C | World Wide Web Consortium |
| Wg | Westgermanisch |

**Box 0.0.1 Hinweise zur Terminologie**

Diese Arbeit steht in einem informatischen und datenwissenschaftlichen Kontext und verwendet damit das dort gebräuchliche Vokabular. Dieses Vokabular ist für gewöhnlich Englisch. Deswegen sind eingeführte Begriffe und Fachtermini weitestgehend in Englisch gehalten. Mit einer Ausnahme: die Begriffe Feature, Merkmal und Eigenschaft werden gleichbedeutend verwendet, wobei der letzte am häufigsten gebraucht wird. Begriffe, die im unmittelbaren Kontext zu der Arbeit stehen oder für diese Arbeit definiert wurden, sind *kursiv* geschrieben. Konzepte, auf die sich in dieser Arbeit bezogen wird, werden durch Kapitälchen markiert. Dies betrifft vor allen die Sprachräume. Auch ist zu beachten, dass bei der Notation von Dezimalzahlen die Vor- und Nachkommastellen durch einen».«und nicht wie im Deutschen üblich mit einem »,«getrennt werden. Dies liegt daran, dass die verwendeten Programme ohne Modifikation diese Art der Eingabe erwarten und auch Ergebnisse in der Form liefern.

# 0 EINLEITUNG

Diese Arbeit behandelt eine Analyse sprachhistorischer Daten mittels Methoden des maschinellen Lernens und die Anwendung einer phonetischen Ontologie als Vorverarbeitungsschritt. Als Anwendungsfall dient dabei der *Mittelrheinischer Sprachatlas* (MRhSA) (vgl. Bellmann, Herrgen und Schmidt 1994–2002), der auf der Onlineplattform REDE[1] (vgl. Schmidt, Herrgen und Kehrein 2008b) als digitaler Atlas abrufbar ist und damit Daten in Form von indizierten Datenbankeinträgen zur Verfügung stellt.

Für die Analyse werden mittels Ontologien die Daten in eine Datenstruktur überführt, die eine Clusteranalyse auf Basis von Lauteigenschaften nach dem IPA Standard ermöglicht. Die dadurch entstehenden Sprachcluster werden anhand objektiver Metriken bewertet und in einen historischen Kontext gesetzt. Da die Daten des MRhSA zwei Generationen umfassen, ist es möglich zeitliche Änderungen hervorzuheben.

Inspiriert wurde die Arbeit von Lameli (2013), der in seiner Habilitationsschrift *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland* eine Reindizierung der deutschen Dialektlandschaft mit vergleichbaren Methoden vornahm.

## 0.1 MOTIVATION

Die Untersuchung sprachlicher Varietäten und deren Abbildung in der Fläche[2] ist spätestens seit der Wenkererhebung (vgl. Wenker 1877; Wenker und Wrede 1888–1923) ein wichtiger Gegenstand der Sprachforschung. Da Sprachwandel ein kontinuierlicher, dynamischer und vor allem langfristiger Prozess ist, ergeben sich für eine Analyse mittels computergestützter Methoden neue Herausforderungen. Viele der Sprachatlanten stammen aus einer Zeit, in der der Einsatz eines Computers entweder beschränkt oder überhaupt noch nicht möglich war. Die Akquise der Daten erfolgte häufig über Jahre; eine anschließende manuelle Nachbearbeitung und Analyse ebenfalls. Die Notation von Lauten ist nicht immer eindeutig oder mit anderen Systemen direkt kompatibel. Auch wenn mit IPA (vgl. International Phonetic Association 1999) ein internationaler Standard existiert, werden doch immer wieder entweder andere Systeme wie Teuthonista bei den *Bayerischen Sprachatlanten* (vgl. Klepsch, Munske und Hinderling 2003; König und Hinderling 1997) oder spezielle Erweiterungen zu IPA verwendet.

Die Erfassung der Daten geschieht nur selten in einem Format, welches ein einfaches, automatisches Auslesen der Kerninformation ermöglicht. Ein weiteres Problem ist die Datenmenge. Auch wenn ein Computer für gewöhnlich sehr effizient bei der Verarbeitung großer Datenmengen sein kann, kann eine Aufbereitung dieser Daten in den meisten Fällen bestenfalls semi-auto-

1 <https://www.regionalsprache.de>, abgerufen 05.02.2018.

2 In dieser Arbeit auch als Sprachraum oder einfach als Raum bezeichnet.

matisch erfolgen und nimmt damit viel Zeit in Anspruch. In Anbetracht der rasanten Entwicklung im Bereich des maschinellen Lernens kann dies dazu führen, dass während der Aufarbeitung der Daten neue Methoden bereitgestellt werden, die allerdings eine andere Aufarbeitungsmethode benötigen. Bei einer Datensammlung über Jahre hinweg fallen bereits Einzelanalysen an. Dies führt zu dem Problem, dass zwar Informationen zu speziellen Aspekten sehr gut verfügbar sind, ein gesamtheitliches Bild aber häufig fehlt oder sich erst sehr langsam herausbildet.

Diese Arbeit versucht diese Probleme anzugehen. Dafür wird zum einen eine Methode zur Normalisierung der erhobenen Lautinformation, mittels einer Ontologie, vorgestellt. Diese Methodik stammt aus dem Kontext der Wissensrepräsentation und kommt besonders bei der Modellierung komplexer, heterogener und verteilter Datenstrukturen zum Einsatz. Ontologien werden bereits in Feldern wie der Medizin, Biologie oder dem Semantischen Web erfolgreich eingesetzt und dienen dort als Grundlage zur Vernetzung oder Strukturierung der anfallenden Daten und Entdeckungen.

Zum anderen ermöglicht eine Clusteranalyse die Verarbeitung aller zur Verfügung stehenden Daten. Das erlaubt eine Raumeinteilung der für den Menschen nicht greifbaren Struktur der Daten, unabhängig von Isoglossen, außerdem kann die Analyse als Bewertung eben dieser gelten und damit als zusätzliche Validierungsmetrik für eine Sprachraumanalyse dienen.

## 0.2 STAND DER DIALEKTOMETRIE

Das Thema dieser Arbeit fällt in den Bereich der Dialektometrie, einem Teilgebiet der Dialektologie, welche sich mit der quantitativen Analyse dialektaler Daten beschäftigt. Wie bereits in Abschnitt 0.1 erwähnt, gibt es eine gewisse Asynchronität zwischen der Erfassung linguistischer Daten und den Möglichkeiten der Analyse. Umfassende quantitative Analysen sind erst durch angemessene Rechenleistung durchführbar geworden, wohingegen die Datenerfassung weniger dadurch begrenzt war. Dies führt dazu, dass dialektometrische Studien in der Regel auf abgeschlossenen Atlasprojekte aufbauen und nicht bereits Teil der Projekte waren. Dies gilt auch für die vorliegende Arbeit. Aktuellere Atlasprojekte ziehen dialektometrische Untersuchungen inzwischen mit in Betracht (zum Beispiel Budin u. a. 2017; Spiekermann u. a. 2016).

Im Folgenden wird eine kurze Übersicht über die Dialektometrie gegeben. Dabei wird der Gegenstand von einer informatischen Sicht beleuchtet und ein besonderes Augenmerk auf die verwendeten Methoden und Prinzipien gelegt (siehe auch Tabelle 0.1).

Als Wegbereiter der Dialektometrie gelten Jean Séguy mit seinem Atlas *Atlas linguistique de la Gascogne* (vgl. Séguy 1973) und Hans Goebl mit seinen *Dialektometrischen Studien* (vgl. Goebl 1984) zu den Sprachtalanten *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS) (vgl. Jaberg, Jud und Scheuermeier 1928) und *Atlas Linguistique de la France* (ALF) (vgl. Gilliéron und Edmont 1902). Die Methodik wird in dem Begleitband *Dialektometrie* (vgl. Goebl 1982) detailliert beschrieben. Dieser Band kann als ein Ausgangspunkt

Tabelle 0.1: Übersicht über die verschiedenen Anwendungen in der Dialektometrie.

| Projekt-standort | Daten-grundlage | Daten-aufbereitung | Analyse-Methode |
|---|---|---|---|
| Salzburg (siehe Goebl 1982) | ALF, AIS | kategorische Daten | RIW, hierarchisches Clustering, Wabenkarten |
| Mainz (siehe Hummel 1993) | Wenkerbögen | kategorische und hierarchische Daten | RIW, Gitternetzkarten |
| Groningen (siehe Heeringa 2004) | u.a. LAMSAS , DNF, SAND | Zeichenketten, kategorische Daten | Levenstein-Distanz, RIW, MDS |
| Salzburg, Ulm (siehe Pröll 2015) | SBS | kategorische Daten, Frequenz-Daten | Kernel-Density-Estimation, K-Means-Clustering, Faktoranalyse, Heatmaps |
| Marburg (siehe Lameli 2013) | Wenkerbögen | Zeichenketten, kategorische Daten, Frequenz-Daten | Split-Trees, Levenstein-Distanz, RIW |
| Freiburg (siehe Szmrecsanyi 2012) | FRED | Ausgewählte Merkmale aus Transkriptionen, Frequenzdaten | Euklidische Distanz, Clustering, MDS, geografische Korrelation |

für die Dialektometrie gesehen werden, insbesondere was die Methodik betrifft.

In der Einführung zu *Dialektometrie* beschreibt Goebl die Problematik der Klassifizierung von (linguistischen) Daten und bedient sich dabei ontologischer Ansatzpunkte (siehe Kapitel 1 dieser Arbeit), ohne den Begriff der Ontologie explizit zu erwähnen. So wird zum Beispiel das Universalienproblem erwähnt und die Taxonomie von Carl von Linné (vgl. Linné und Gmelin 1788) als Beispiel für eine Klassifikation präsentiert. Der Gegenstand, der in dieser Arbeit als Datenset bezeichnet wird, wird in den Arbeiten von Goebl in Anlehnung an diese historische Taxonomie Taxandum genannt.

Als Grundlage dieser Taxate dienen die kartierten Sprachdaten des AIS und des ALF. Dabei wurden die Datensets erzeugt, indem das Kartenthema als Merkmal beziehungsweise Variable und die entsprechenden Ausprägun-

gen oder Realisationen des Kartenthemas an den Belegorten als die Merkmals- oder Variablenausprägungen aufgefasst wurden. So lassen sich Datensets auf Basis der Kartenthemen erstellen. Für die *Dialektometischen Studien* wurden so verschiedene Taxate erstellt:

1. Ein vollständiges Datenset, in dem alle Karten aufgeführt wurden
2. Ein Datenset, das nur Karten enthält, die zu jedem Ortspunkt Merkmalausprägungen haben.
3. Teildatensets, die jeweils nur eine Auswahl an Karten enthalten. Diese Teildatenset sind aber nicht notwendigerweise distinkt.

In Tabelle 0.2 ist der schematische Aufbau der Taxate dargestellt.

Tabelle 0.2: Schematischer Aufbau der Taxate von Goebl. Jedem Kartenthema und jeden Belegort ist eine entsprechende Realisation zugeordnet.

| BELEGORT | KARTENTHEMA I | KARTENTHEMA II | ... |
|---|---|---|---|
| Ort 1 | Realisation 1A | Realisation 2A | ... |
| Ort 2 | Realisation 1B | Realisation 2B | ... |
| Ort 3 | Realisation 1C | Realisation 2C | ... |
| | ⋮ | | |

So können zu jedem Ort im Untersuchungsgebiet sogenannte Merkmalsstränge konstruiert werden, die zu jedem Ort die entsprechenden Merkmalausprägung jedes Kartenthemas beinhalten. Goebl vertrat die Annahme, dass größere Merkmalstränge zu valideren Ergebnissen führen, eine Annahme, die heute nicht mehr so vertreten wird (siehe dazu Abschnitt 3.2). Um zu ermitteln, wie ähnlich sich die Dialekte zweier Orte sind, können nun die Merkmalsstränge miteinander verglichen werden. Zur Berechnung der Ähnlichkeit zweier Merkmalsstränge verwendet Goebl eine Metrik, die er als *Relativer Identitätswert* (RIW) bezeichnet:

$$\text{RIW} = 100 * \frac{\#\text{Übereinstimmende Merkmale}}{\#\text{Alle Merkmale}}$$

Diese Formel ist äquivalent zu dem Jaccard-Koeffizenten (vgl. Jaccard 1912) zur Ähnlichkeitsbestimmung zweier Mengen[3]. Ursprünglich für die Botanik entwickelt, gilt der Jaccard-Koeffizent als eine der Basismethoden zur Ähnlichkeitsbestimmung von Texten (vgl. Huang 2008) und ist deshalb auf dem Gebiet des „Text-Mining“ sehr prominent. Mithilfe dieser Formel lässt sich eine Ähnlichkeitsmatrix der Merkmalspunkte zueinander aufstellen, die jedem Ortspunktpaar einen prozentualen Ähnlichkeitswert zuweist. Da die Merkmalsstränge nicht notwendigerweise vollbesetzt sein müssen[4], werden

3 Die nominalen Merkmalstränge von Goebl können als Mengen aufgefasst werden.
4 Zum Beispiel wenn ein Ort zu einem Kartenthema nicht befragt wurde.

Spalten, die mindestens einen leeren Wert enthalten bei der Berechnung des RIW ignoriert[5].

Eine Spalte (oder Zeile) dieser Distanzmatrix kann nun als Choroplethenkarte visualisiert werden. Dazu wird ein Referenz-Ort[6] ausgewählt und die Distanz (definiert als : 1-RIW) zu den übrigen Orten[7] gilt als Darstellungsgrundlage. Zur besseren Übersicht wurden die Karten in sechs beziehungsweise zwölf Klassen eingeteilt. Diese Technik wird auch *Binning*[8] genannt. Zur Einteilung in diese Klassen werden drei Methoden verwendet. Die erste Methode unterteilt das Spektrum vom Minimum bis zum Mittelwert der Daten in $n$ gleiche Intervalle[9] und dann nochmal vom Mittelwert bis zum Maximum der Daten. Sie wird als *MINMWMAX* bezeichnet und verwendet ein $n$ von sechs. Die zweite Methode, die als *MEDMW* bezeichnet wird, funktioniert ähnlich, nur dass – anstelle einer Unterteilung in Intervalle gleicher Länge – jede Klasse die gleiche Anzahl von Datenpunkten enthält. Die letzte Methode unterteilt das ganze Spektrum vom Minimum bis zum Maximum in Klassen, so dass jede Klasse die gleiche Anzahl von Ortspunkten enthält. Sie wird *MED* genannt. Eine Visualisierung der Choroplethenkarte erfolgt entweder mittels einer diskreten Farbskala oder eines schwarz-weißen Musters zur Unterscheidung der einzelnen Klassen. Zusätzlich zu den aufgearbeiteten Choroplethenkarten liegen die Rohkarten als ASCII-Karten[10] vor.

Diese Form der Visualisierung zeigt nur die Dialektähnlichkeiten zu einem ausgewählten Ort, für einen besseren Vergleich der strukturellen Ähnlichkeiten der Orte untereinander benötigt es andere Methoden. Eine Kartenart, die dabei helfen kann, sind die Wabenkarten. Dabei werden die Orte wieder als Polygone modelliert, die durch die Voronoi-Tessellation ein wabenförmiges Muster annehmen. Die Grenzen zwischen zwei benachbarten Orten werden je nach errechneter Ähnlichkeit unterschiedlich dick eingezeichnet. Diese Methode bietet eine schnelle Möglichkeit deutliche dialektale Unterschiede zwischen zwei Gebieten zu visualisieren und fungiert gewissermaßen als eine Art datenbasierter Isoglossen-Generator. So kann man zum Beispiel einen Schwellenwert definieren und nur Grenzen über diesem Schwellenwert anzeigen lassen.

Neben den Distanzmatrizen wurden auch bereits Clusteringverfahren angewendet, meistens in der Form von hierarchischem Clustering und insbesondere in Form von Dendrogrammen (siehe dazu Abschnitt 3.3). Allerdings

5 Eine andere Möglichkeit wäre das Fehlen einer Variante mittels eines speziellen Symbols zu repräsentieren oder dem mengentheoretischen Hintergrund des Jaccard-Koeffizienten folgend keine speziellen Anpassungen vornehmen und einfach mit der Vereinigung bzw. dem Schnitt beider Mengen arbeiten.

6 Für gewöhnlich ein Ort, der dialektal nah an dem gewählten Standard angesehen wird.

7 In den Karten werden die Orte als Polygone dargestellt. Eine Karte ist also nicht eine Ansammlung von Ortspunkten, sondern unterteilt das Untersuchungsgebiet in Polygone, wie es bei Choroplethenkarte üblich ist. Für die Erstellung von Polygonen aus Punkten bedient man sich für gewöhnlich der Voronoi-Tessellation (vgl. Aurenhammer und Klein 2000).

8 Das *Binning* ist eine Datenglättungstechnik, um ein kontinuierliches Spektrum in diskrete Klassen zu unterteilen.

9 Die Intervalle entsprechen damit den Klassen.

10 Eine Methode aus den Anfängen der Computergrafik, bei der Karteninformationen mittels Tastaturzeichen dargestellt werden. So können zum Beispiel Grenzen mittels einer Folge von $- oder %-Zeichen repräsentiert werden.

wird keine vollständige Clusteranalyse durchgeführt, sondern die Dendrogramme dienen zum Hervorheben besonderer Eigenschaften des Taxats.

Goebls Methodik und der relative Identitätswert sind bis heute ein Bestandteil der Dialektometrie. Zusätzlich zu dem RIW führte Goebl den gewichteten RIW (vgl. Goebl 1984) ein, um ein Datenset besser in Bezug auf seltene Phänomene kontrollieren zu können, dabei wird ein Merkmal mit einem Faktor multipliziert, der von der Häufigkeit des Merkmals im Taxat abhängig ist. Der RIW, als prominentes Distanzmaß in der Dialektometrie, wird in vielen dialektometrischen Studien angewendet, dabei wird allerdings nur selten die Verbindung zu dem Jaccard-Koeffizent oder dem damit zusammenhängenden „Text-Mining" herausgestellt (vgl. auch Lameli 2013, S. 49). Der Ansatz von Goebl eignet sich besonders für Datensets, die für jedes Kartenthema jedem Ortspunkt genau eine Merkmalsausprägung zuweisen.

Direkt anschließend an die Analysen von Goebl bietet der *Kleine Deutsche Sprachatlas* (KDSA) (vgl. Veith, Hummel und Putschke 1984–1999) mit den dazugehörigen *Dialektometrischen Studien* (vgl. Hummel 1993) einen weiteren Atlas, der sich insbesondere dadurch auszeichnet, dass er von vornherein als computergestützter Atlas geplant war, inklusiver eines – für die Zeit – professionellen Datenbanksystems für die Modellierung und Speicherung der Daten. Als Datengrundlage dient dabei eine repräsentative Auswahl (vgl. Veith 1984, S. 304) der Wenkererhebung (vgl. Schmidt und Herrgen 2011, S. 97–107) an 5892 der ca. 50000 Befragungsorte. Da die Wenkerbögen nicht direkt übernommen werden können, wurden die Daten einer umfassenden Vorverarbeitung unterzogen, um die in den Sätzen enthaltenen linguistische Phänomene besser fassen zu können. Aus den Wenkersätzen wurden dazu 174 Morphe ausgewählt, das heißt, Stämme und Affixe. Diese wurden weiter in ihre standardsprachlichen Tagmen zerlegt. Diese Tagmen[11] wurden entsprechend ihrer Artikulationsmerkmale klassifiziert (vgl. Veith 1984) und bilden die Grundlage für die 462 Kartenthemen. Diese Art der Klassifikation kann dazu führen, dass eine Realisationsform je nach Kontext mehreren Morphen oder Tagmen zugeordnet ist. Um den Kontext zu behalten, wurden die Ausprägungen an den Belegorten von einer kategorischen Struktur in eine binäre[12] überführt. Dafür wird eine Technik namens One-Hot-Encoding (siehe Abschnitt 3.2) angewendet, allerdings ohne diese Technik explizit zu benennen[13]. Für die dialektometrische Auswertung wurden verschiedene Subsets dieser Kartenthemen wie zum Beispiel Frikative oder Langvokale generiert, die als Basis für die 266 erstellten Karten dienen. Die Auswertung selbst ähnelt der von Goebl. Wieder kommen der RIW in Relation zu einer „Standardlautung" oder einem ausgewählten Ortspunkt und eine Klassifizierung anhand der Verteilungsspektren (zum Beispiel MED) zum Einsatz. Eine Kartierung erfolgt sowohl thematisch als auch regional auf Basis eines quadratischen Gitternetzes. Die Visualisierung erfolgt anhand von deskriptiven Piktogrammen in Schwarz-Weiß. Nachbarschaftsbeziehungen werden wieder an Hand der Liniendicke zu den vier direkten Nachbarn eines Quadrates

11 Bei Hummel (1993) werden diese Tagmen in Anlehnung an Goebl auch als Taxate bezeichnet.

12 Es spielten auch praktische Gründe auf Grund der damals zur verfügung stehenden Rechenleistung eine Rolle.

13 Der Term hat sich erst später mit dem großflächigen Einsatz des Computers in der Datenanalyse durchgesetzt.

visualisiert, diese sind allerdings auf Grund der schieren Menge sehr unübersichtlich und lassen nur großflächige Muster erkennen.

Der KDSA bietet ein sehr gutes Beispiel für den frühen Einsatz computergestützter Analysen und der Datenvorverarbeitung. Durch die Vorverarbeitung sind die Daten nach phonetischen Merkmalen vorsortiert und erlauben so die Erstellung von Karten unter bestimmten Gesichtspunkten, wie zum Beispiel eine Karte zu allen Langvokalen oder Plosiven.

In den 90er- und den frühen 2000er-Jahren gab es einen Boom an Entwicklungen in dem Bereich des „Information Retrieval" und der Biologie. Dank leistungsfähiger Hardware konnten nun viele Algorithmen auf Probleme oder Datensets angewandt werden, die vorher nicht in akzeptabler Zeit analysiert werden konnten. Dies führte zum einen zu dem neuen Forschungsfeld der Bioinformatik, aber auch zu vielen neuen Methoden in der Textanalyse und der Datenanalyse generell.

Bereits früh wurde die Multidimensionale Skalierung (MDS) (siehe auch Abschnitt 3.4) für die Dialektometrie entdeckt (vgl. Embleton 1993). MDS kann eine Antwort auf die Frage bieten, wie man die Distanzen zwischen allen Orten sichtbar machen und gleichzeitig ein interpretierbares Kartenbild wahren kann. Goebl umging diese Problematik, indem er einen Referenz-Ort auswählte. Alle Karteneinfärbungen stehen also in Abhängigkeit von diesem Ort. In der Wabendarstellung werden nur die dialektalen Distanzen zwischen benachbarten Orten betrachtet.

MDS ist ein sogenanntes dimensionenreduzierendes Verfahren und dient dazu, einen hochdimensionalen Raum wie zum Beispiel eine Datenmatrix von dialektalen Distanzen in einen niederdimensionalen Raum zu überführen, ohne dabei die relative „Struktur" der einzelnen Datenpunkte untereinander zu stark zu verändern. Für gewöhnlich wird auf drei Dimensionen reduziert, da man mit drei Komponenten Farben am Computer generieren kann. So können Choroplethenkarten generiert werden, bei denen eine Farbähnlichkeit einer dialektalen Ähnlichkeit entspricht. Dies ermöglicht auf einfache Art und Weise ein stetiges Raumbild dialektaler Strukturen zu generieren. Die Ergebnisse einer MDS lassen sich bewerten, so kann zum Beispiel die „goodness of fit" mittels des Bestimmtheitsmaßes ($R^2$) bestimmt werden. Dieses Maß gibt an, wieviel Varianz die reduzierten Dimensionen im Verhältnis zu den originalen Dimensionen abbilden können. Je höher $R^2$ ist, umso besser spiegelt also eine Choroplethenkarte die dahinter liegende Ausgangsdatenstruktur wieder. MDS ist ein exzellentes Werkzeug, um schnell komplexe Daten zu visualisieren, allerdings benötigt es eine entsprechende Datengrundlage in Form einer Distanz- oder Ähnlichkeitsmatrix. Auch muss beachtet werden, dass diese Art der Visualisierung nur eine Annäherung bis zu einem bestimmten Grad sein kann und gerade komplexe Datenstrukturen durchaus verzerrt sein können.

Eine andere Methode zur Distanzberechnung, die besonders in der Bioinformatik Verwendung findet, aber generell für den Vergleich von Zeichenketten geeignet ist, ist die Levensthein-Distanz (vgl. Kruskal 1983; Levenshtein 1966). In der Biologie werden DNA oder RNA-Sequenzen als Zeichenketten, die aus dem Alphabet A-T/U-G-C bestehen, kodiert. Ziel ist es, die Ähnlichkeit zweier RNA-Squenzen zu bestimmen. Diese Sequenzen sind sehr lang, allerdings bestehen sie nur aus dem oben genannten Alphabet.

Trotzdem wird für einen Vergleich so viel Rechenleistung benötigt, dass eine Anwendung erst seit den 90er-Jahren praktikabel ist. In der Linguistik kann die Levensthein-Distanz ähnlich verwendet werden, nur sind die Wortsequenzen für gewöhnlich kürzer, allerdings das Alphabet umfangreicher. So kann die Ähnlichkeitsbestimmung zwischen zwei Wörten oder Sätzen, wie zum Beispiel auf Basis transliterierter Wenkerbögen (vgl. Nerbonne und Siedle 2005) vorgenommen werden und als Alphabet entweder der Standard-Zeichensatz (zum Beispiel das lateinische Alphabet) oder auch ein phonetisches Alphabet dienen. Die Levensthein-Distanz ergibt sich aus der minimalen Anzahl der Operationen *Keine Änderung*, *Ersetzen*, *Löschen*, die benötigt werden, um eine Zeichenkette in eine andere zu überführen. Dabei können die Operationen gewichtet werden. Meistens wird für *Keine Änderung* 0 gesetzt und 1 für die anderen beiden[14]. Die Levensthein-Distanz wird mittels eines dynamischen Algorithmus bestimmt. Dabei wird versucht, das komplexe Gesamtproblem in weniger komplexe Teilproblem zu unterteilen, diese zu lösen und anschließend mittels einer Technik namens *Backtracking* eine optimale Lösung für das Hauptproblem zu finden. Hierbei werden die beiden Zeichenketten in Zeilen-Spalten-Form gegenüber gestellt und iterativ für jede Zelle der Modifikationswert berechnet. Dieser Wert setzt sich aus der aktuellen Operation plus dem Minimum der benachbarten vorhergehenden Operationen zusammen. Als mathematische Formel sieht der Algorithmus zur Berechnung der Levensthein-Distanz (ld) so aus:

$$\mathrm{ld}_{a,b}(i,j) = \begin{cases} \max(i,j), \text{wenn } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{ld}_{a,b}(i-1,j) + 1 \\ \mathrm{ld}_{a,b}(i,j-1) + 1 \\ \mathrm{ld}_{a,b}(i-1,j-1) + \mathrm{op}_{(a_i \neq b_j)} \end{cases} & \text{sonst} \end{cases}$$

Wobei gilt, dass $a$ und $b$ je zwei Zeichenketten sind, $i$,$j$ die Positionen in den Zeichenketten und $\mathrm{op}_{(a_i \neq b_j)}$ die Modifikationsoperation. Anders als zum Beispiel der RIW, der zwischen 0 und 100 genormt ist, ist die Levensthein-Distanz abhängig von der Länge der zu vergleichenden Zeichenkette. Das macht diese Art der Distanzmessung kontextabhängig und ohne eine Normierung ist der Vergleich zwischen zwei Datensets nicht möglich.

Diese Levenshtein-Distanz als Distanzmetrik wird an der Universität Groningen in der Arbeitsgruppe um John Nerbonne seit den späten 1990ern (vgl. Nerbonne und Heeringa 1997; Nerbonne, Heeringa und Kleiweg 1999)[15] und besonders seit der Dissertation von Wilbert Heeringa (vgl. Heeringa 2004) erfolgreich in vielen Analysen verwendet, so dass sich Groningen neben Salzburg mit Hans Goebl als ein Zentrum für Dialektometrie etabliert hat. So wurde der *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) (vgl. McDavid Jr und O'Cain 1980) einer dialektometrischen Analyse unterzogen, dabei wurde für einfache lexikographische Vergleiche die Methodik

14 Die Gewichtung dieser Operationen ist kontextabhängig und oft nicht eindeutig geklärt. Manchmal wird *Ersetzen* nur mit 0.5 angegeben oder es werden komplexe Kostenfunktionen definiert.

15 Eine erste Anwendung in der Dialektologie fand die Levensthein-Distanz auf den Daten zum *Linguistic Atlas and Survey of Irish Dialects* (vgl. Wagner 1958-1969) in *Computational Dialectology in Irish Gaelic* von Kessler 1995.

von Goebl eingesetzt, für feinere Unterscheidungen die Levensthein-Distanz (vgl. Nerbonne und Kleiweg 2003). Visualisiert wurden die Ergebnisse zum einen mittels Einfärben von erstellten Clustern aber auch mittels Multidimensionaler Skalierung. Mit dieser Methodik wurden im Laufe der Zeit mehrere Dialekträume untersucht. Dabei folgt die Vorgehensweise einem ähnlichen Muster: Mittels Levensthein-Distanz wird eine Distanzmatrix für ein Datenset erstellt und mittels MDS wird diese visualisiert. Als zusätzliche Visualisierung werden auch sogenannte *Netzwerk-Karten* herangezogen. Diese funktionieren ähnlich der Wabenkarten von Goebl und zeigen die Ähnlichkeit zwischen je zwei Ortspunkten mittels Linien zwischen zwei Polygonen oder Ortspunkten an. Dabei wird für gewöhnlich wieder auf besonders relevante Informationen gefiltert, um die Übersicht zu wahren.

Als Datengrundlage für dialektometrische Studien für die Niederlande dienen der *Reeks Nederlandse Dialectatlassen* (RND) (vgl. Blancquaert und Pée 1925–1982), das *Goeman-Taeldeman-Van Reenen-Project* (GTRP) (vgl. Reenen, Goeman und Taeldeman 2003) und der *Syntactic Atlas of the Dutch Dialects* (SAND) (vgl. Barbiers u. a. 2005, 2008 ). Die Hauptarbeit dazu ist *Measuring Dialect Pronunciation Differences using Levenshtein Distance* von Heeringa (2004), die unter anderem den RND einer Analyse unterzieht. Diese Dissertation dient als Ausgangspunkt vieler weiterer Studien. So wurden von Wieling, Heeringa und Nerbonne (2007) die Daten des GTRP und RND verglichen. In *Associations Among Linguistic Levels* präsentieren Spruit, Heeringa und Nerbonne (2009) einen Teilaspekt der Ergebnisse für eine Analyse des SAND. Dabei wird ein Augenmerk auf den Zusammenhang der unterschiedlichen Modalitäten der Datensets gelegt und die Korrelation der Syntax des SAND zu Lexik und Aussprache des RND untersucht und in Form von MDS-Karten präsentiert. Auch wurden bereits Studien zur deutschen Dialektlandschaft mit diesen Methoden vorgenommen. So wurden auf Basis der transkribierten PAD-Daten (vgl. Nerbonne und Siedle 2005) Vokale und vokalisch bedeutsame Worte untersucht und eine Dialekteinteilung auf Basis signifikanter Ausspracheunterschiede vorgenommen (vgl. Nerbonne 2009, S. 178 ff.). Über die Jahre wurde diese Methodik um eine Auseinandersetzung mit der Qualität der Datensets und der Validierung der Ergebnisse erweitert. Ein weiteres Augenmerk liegt auf dem Zusammenhang zwischen dialektaler und geographischer Distanz. Außerdem wird von der Arbeitsgruppe um Nerbonne eine Online-Plattform *Gabmap*[16] (vgl. Nerbonne u. a. 2011) zur Verfügung gestellt, die das Erstellen eigener Karten und Analysen mittels Levensthein-Distanz oder RIW, MDS und Clustering ermöglicht.

Eine andere Art zur Dialekteinteilung, die auch auf Clustering beruht, ist die sogenannte *Split-Tree*-Analyse[17] (vgl. Huson 1998) und als Erweiterung die *Neighbor-Net*-Methode (vgl. Huson und Bryant 2006). Diese Methoden stammen auch aus der Biologie, genauer der Phylogenetik, wo zum Bespiel versucht wird, anhand des genetischen Codes oder der Proteinstrukturen die historischen Verwandschaftsbeziehung zwischen Arten zu bestimmen und einen Stammbaum zu rekonstruieren[18]. Die *Neighborhood-Nets* und *Split-*

16 <https://gabmap.nl>, abgerufen 01.04.2020.

17 *Split-Trees* werden sehr oft mittels der an Universität Tübingen entwickelten gleichnamigen Software erstellt.

18 Auch hier sei wieder Carl von Linné als Wegbereiter erwähnt.

*Trees* werden sehr prominent in der vergleichenden Linguistik zur Klassifizierung von Sprachfamilien eingesetzt (vgl. Nichols und Warnow 2008). In der Dialektologie kann die Arbeit von Lameli (2013) als Hauptwerk angesehen werden. Diese Techniken dienen dort auch als Grundlage seiner integrativen Dialekteinteilung (vgl. Lameli 2019, S. 158 ff.; Schmidt 2017, S. 107). Das besondere an *Split-Trees* ist, dass anders als zum Beispiel bei Dendrogrammen, bei denen eine strikt hierarchische Einteilung der Daten vorgenommen wird, auch die relativen Ähnlichkeiten der Datenpunkte berücksichtigt werden. Die Daten werden also nicht rein von einer sichtbaren Wurzel aus in eine Richtung angeordnet, sondern können in jede Richtung von der Wurzel aus platziert werden. Die Wurzel selbst verliert dabei an Bedeutung, weshalb *Split-Trees* auch als wurzelfreie Bäume bezeichnet werden. So erhält man einen Graphen, in dem die Knoten und Blätter relativ zueinander in einer Ebene angeordnet sind. Dies ermöglicht einen schnellen Überblick über die Verwandschaftsbeziehungen und darüber, wie die verschiedenen Teilcluster in dem Graphen untereinander in Beziehung stehen.

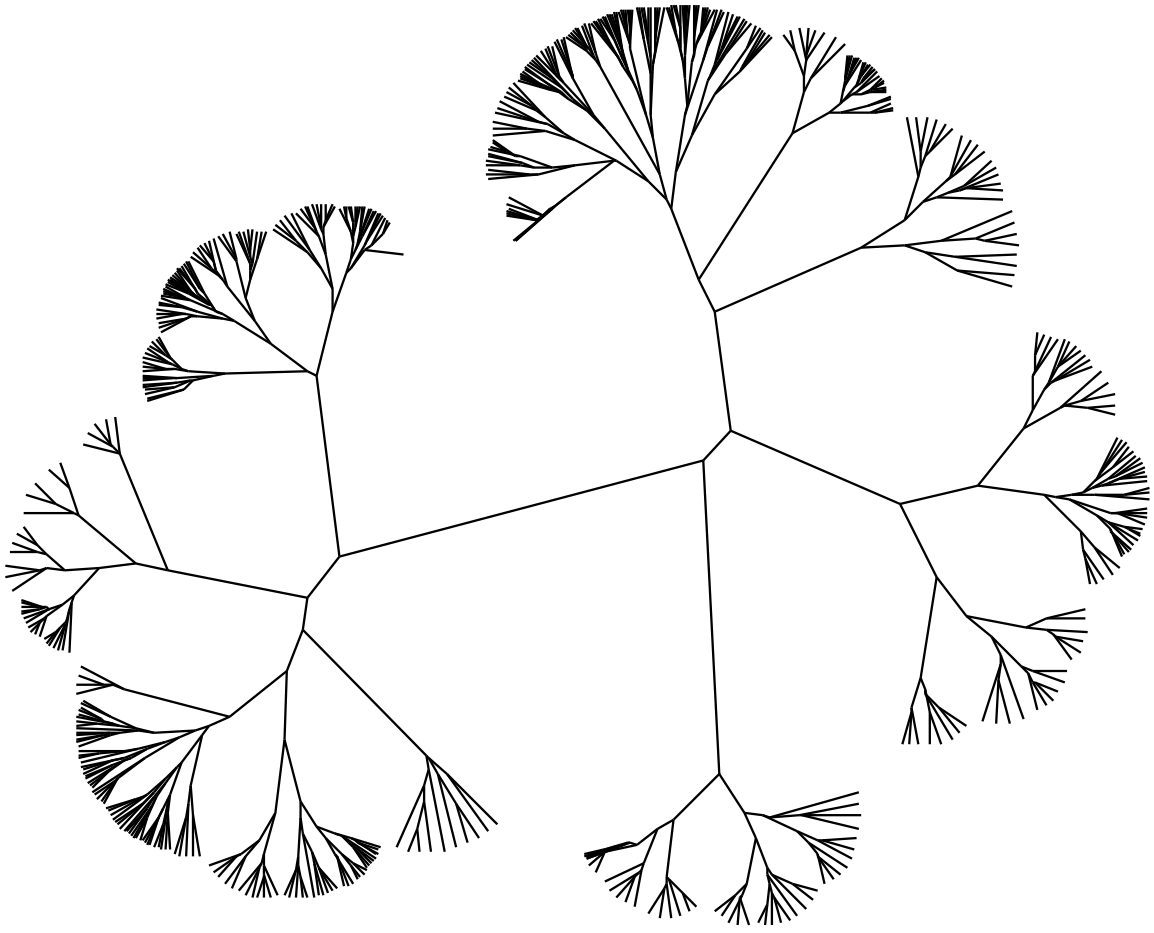


Abbildung 0.1: Beispiel eines *Split-Trees*. Dieser Baum basiert auf Daten dieser Arbeit. Auf eine genaue Kennzeichnung der Blätter wurde bewusst verzichtet.

Die Art, wie der Graph aufgebaut wird, kann sich je nach Struktur der Grunddaten deutlich unterscheiden. Verwendet man eine distinkte Eigenschaftsmenge, wie zum Beispiel bei dem Taxandum von Goebl, ist eine Distanzmetrik nicht unbedingt nötig und der Graph kann direkt auf Basis der Merkmalsausprägungen erstellt werden. Die Erstellung eines *Split-Trees* ist wieder sehr rechenintensiv und oft nicht eindeutig in endlicher Zeit lösbar[19]. Deswegen wird sich dabei des sogenannten *Bootstrappings* bedient. Das ist eine Technik, bei der die Struktur eines Datensets bestimmt wird, indem die Operationen zur Generierung des Baums auf zufällig ausgewählten Teilmengen wiederholt ausgeführt werden und sich zum Schluss ein Gesamtergebnis auf Basis der Auswahlwahrscheinlichkeiten bildet (mehr dazu in Abschnitt 3.4). Da Bootstrapping auch als eine Verifikationstechnik angesehen werden kann, bieten *Split-Trees* ein sehr mächtiges Werkzeug für die Clusteranalyse dialektaler Daten. Die *Neighbor-Nets* sind eine Erweiterung der *Split-Trees* und zeichnen sich durch eine genauere Aufschlüsselung der relativen Verwandtschaft aus. Während *Split-Trees* nur einen Pfad zwischen zwei Blättern oder Knoten erlauben, können in den *Neighbor-Nets* auch zusätzliche direkte Verbindungen angezeigt werden. Dabei kann die Dicke der Verbindungen wieder als Metrik für die Ähnlichkeit angesehen werden.

Einen etwas anderen Fokus in der Dialektometrie legt das Gemeinschaftsprojekt *Geoling*[20] der Universitäten Augsburg und Ulm. Während viele Methoden sich mit der Zuordnung und Klassifizierung von größräumigen Dialektgebieten beschäftigen, wird bei diesem Projekt der Fokus auf die internen Strukturen eines Untersuchungsgebiets gelegt. Untersucht wird die interne, räumliche Struktur und besonders, welche latenten Eigenschaften diese Sprachregionen auszeichnen. Die Datengrundlage bildet der *Sprachatlas von Bayerisch-Schwaben* (vgl. König 1996–2006) und die Dissertation *Raumvariation zwischen Muster und Zufall: Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben* von Pröll (2015) kann als das Hauptwerk dazu angesehen werden. Für die Teilaspekte dieser Art der Analysen sind in Einzelpublikationen noch einmal gesondert erläutert. Dabei wird neben einer diskreten Abgrenzung der einzelnen Sprachräume in Form einer Clusteranalyse auch Wert auf eine Visualisierung des Datenspektrums in Form von Gradienten gelegt (vgl. Rumpf u. a. 2010). Dabei wird sich Methoden aus der statistischen Geographie und der statistischen Bildanalyse bedient. Die untersuchten Orte wurden wieder mittels Voronoi-Tessellation in Polygone umgewandelt, so dass sich auf Basis dieser Polygone ein stetiges Bild ergibt. Mit Hilfe sogenannter *Kernel-Density-Estimation* (vgl. Parzen 1962; Silverman 1986) wird dabei ein statistisches Modell für eine Ausprägungswahrscheinlichkeit einer Variante[21] in einem Polygon erstellt. Durch die *Kernel-Density-Estimation* wird die eine abgeschwächte Ausprägungswahrscheinlichkeit auf Basis eines Kernels auch auf die nahen Polygone übertragen[22].

19 Das Erstellen eines *Split-Trees* ist ähnlich wie die Levensthein-Distanz ein Optimierungsproblem mit Backtracking, nur noch komplexer, da die Verarbeitungsmöglichkeiten der Datenpunkte nicht durch die vorhergegangen limitiert werden.

20 <https://www.geoling.net>, abgerufen 05.04.2020.

21 Eine Variante kann in diesem Fall nicht nur eine linguistische Variante sein, sondern auch das Ergebnis einer statistischen Berechnung (zum Beispiel eines Clusterings).

22 Das kann man sich ähnlich einer Glühbirne vorstellen, bei der Licht abgestrahlt wird und sich nach und nach abschwächt. Wenn zwei Lichtquellen nah bei einander sind, überlagern

So können sich zum Beispiel Übergangsgebiete bilden, in denen zwei Cluster gegenseitig in den anderen „hineinstrahlen“. Diese Methode ist in der statistischen Geographie beliebt, da man dort statistische Vermutungen über eine Region, in der man keine Daten erheben kann, auf Basis der Nachbarschaft anstellen kann. So lassen sich Flächenkarten erzeugen, bei denen jede Region durch den errechneten Farbwert der dominierenden Variante repräsentiert wird. Da bei den meisten Karten eine Einfärbung allerdings an die Polygone gebunden ist, tritt der Funktionen-Charakter des Kernels in den Hintergrund und die Einfärbung folgt eher einer Art des *Binnings*.

Eine Besonderheit dieser Untersuchung ist, dass nicht versucht wird, eine räumliche Struktur auf Basis aller Daten zu finden, sondern es wird eine Ähnlichkeit der Karten auf Basis ihrer räumlichen Strukturen ermittelt. Dieses Vorgehen erlaubt es, unterschiedliche Datenstrukturen zu vergleichen und bietet Möglichkeiten nicht nur räumliche, sondern auch thematische Zusammenhänge zu untersuchen. Dies bietet sich an, da die Karten des SBS auch thematisch kategorisiert sind. Für eine Untersuchung der Karten untereinander wird auf Basis der Raumeinteilung innerhalb der Karten eine globale Distanz zwischen je zwei Karten generiert. Aus diesen Distanzen wird anschließend eine Distanzmatrix für alle Karten generiert und mit Hilfe hierarchischen Clusterings geclustert. Anschließend wird mit Hilfe statistischer Tests überprüft, ob und welche Kategorien mit den Clustern signifikant korrelieren und wie die Karten eines Clusters untereinander in Beziehung stehen.

Neben der Klassifizierung der Karten auf Basis einer repräsentativen Raumstruktur wurde auf den Daten des SBS eine sogenannte Faktor-Analyse durchgeführt (vgl. Pröll, Pickl und Spettl 2014). Eine Faktor-Analyse versucht sogenannte latente Variablen als Kombination eines dritten Faktors zu beschreiben. Ein Ziel ist es, zu bestimmen, in wieviele voneinander unabhängige Faktoren sich die Daten zerlegen lassen. Dies ist ähnlich wie die MDS eine Art der Dimensionsreduktion. Die Stärke der Korrelation einer latenten Variable mit einem Faktor kann außerdem Aufschluss über die Bedeutung dieser Variable liefern. Auf die Daten des SBS angewandt kann nun die Verteilung der einflussreichsten Faktoren und den dahinterliegenden Variablen bestimmt werden. So kann man zeigen, welche Phänomen in welchen Regionen besonders dominant sind. Gruppiert man die Regionen nach den wichtigsten Faktoren, kann diese Methode zu einer Dialektraumeinteilung führen.

Die vorgestellten Methoden und Projekte geben eine Übersicht über die Anwendung der Datenwissenschaft in der Dialektometrie wieder. Gerade aus der Bioinformatik haben sich viele Methoden etabliert und dank MDS gibt es eine Möglichkeit, leicht Daten zu visualisieren. Allerdings sollte man beachten, dass die Dialektometrie ein überschaubares Feld ist und häufig auf bereits etablierte Methoden zurückgreift. Als Datengrundlage dienen in der Regel Sprachatlanten, deren Datenstruktur nicht für eine großflächige, computergestütze Auswertung angelegt wurde.

Auch wenn hier von Sprachatlanten als Ausgangsbasis gesprochen wird, ist eine quantitative Analyse nicht auf Kartendaten beschränkt. Natürlich

sich diese Lichtquellen und die Strahlen erscheinen heller. Im statistischen Sinne versucht eine *Kernel-Density-Estimation* eine Verteilungskurve auf Basis von einer Stichprobe zu berechnen.

kann auch ein (Daten-) Korpus als Basis dienen, wie zum Beispiel der *Freiburg Corpus of English Dialects* (FRED) (vgl. Szmrecsanyi und Hernández 2007). Die Analysen von Szmrecsanyi 2012 beziehen sich auf die Daten zu den 34 Counties, die im Rahmen des FRED-Projekts erhoben wurden[23]. Er verwendet dort unter anderem die Häufigkeitsverteilungen dialektaler Varianten für ausgewählte sprachliche Merkmale. Dabei verwendet er als Distanzmaß die Euklidische Distanz und liefert neben einer MDS auch Clusterings zu den Daten.

Ein interessanter Aspekt dieser Arbeit ist die Korrelation mit dritten Faktoren, in diesem Fall die geografische Distanzen zwischen den Orten. Diese Art von Vergleich erfreut sich gerade im Zusammenhang mit der Soziolinguistik einer gewissen Beliebtheit in der Dialektomertrie (siehe auch Gooskens 2004; Nerbonne 2007; Trudgill 1974 und in Korrelation mit wirtschaftlichen Aspekten bei Falck, Lameli und Ruhose 2018).

Für hochwertige Analysen ist die Datenvorverarbeitung mindestens genauso wichtig wie die Analyse selbst. Auch sollten Analysen einer Überprüfung der Ergebnisse unterzogen werden, wobei es nicht ausreicht, diese Ergebnisse mit externen Annahmen zu vergleichen, sondern auch die interne Struktur der Daten und die interne Struktur der Ergebnisse sollten häufiger für die Verifikation herangezogen werden.

Diese Arbeit legt den Fokus auf diese Bereiche. In Kapitel 1 wird eine Methode zur Strukturierung und Aufarbeitung von Daten besprochen und anschließend in Kapitel 2 auf die Daten des MRhSA angewendet. In Kapitel 3 werden die Methoden zur Clusteranalyse vorgestellt und in Kapitel 4 wird das Datenset mittels dieser Methoden überprüft und verifiziert.

23 Die Datenbasis besteht aus 431 Informanten in insgesamt 156 Orten.

# Teil I

# ONTOLOGIEN UND DATEN

# 1

# EINE ONTOLOGIE FÜR DIE PHONETIK

Dieses Kapitel bietet eine Einführung in die Wissenschaft der computergestützten Wissensrepräsentation. Neben einer kurzen Erläuterung der häufig verwendeten Terme und einem kurzen historischen Überblick, beschreibt es die verwendeten Techniken und dahinter liegenden Prinzipien zur Ontologieerstellung. Es schließt ab mit einer Beschreibung der *phonOntology*, der für diese Arbeit entwickelten Ontologie.

## 1.1 WISSENSREPRÄSENTATION

Wissensrepräsentation ist ein zentrales Thema der Informatik und beschäftigt sich mit der Beschreibung von Daten, um einen semantisch interpretierbaren Mehrwert in Form von Wissen[24] zu erhalten. Der Computer dient dabei als Werkzeug zur Erzeugung dieser Strukturierung und als Speichermedium. Zudem stellt er Methoden für einen effizienten Zugriff auf die Daten und zur Weiterverarbeitung der Daten zur Verfügung. Wissensrepräsentation im Computer ist zudem eine Projektion eines Ausschnittes[25] der Welt auf eine durch den Computer interpretierbare Struktur mithilfe einer formalen Sprache[26]. Das Modellieren[27] dieser Wissensrepräsentation ist für gewöhnlich zielgerichtet. So kann sich die Repräsentation eines Gespräches sehr unterscheiden, je nachdem ob man es unter einem soziologischen Gesichtspunkt[28] oder unter einem biologischen Gesichtspunkt modelliert. So kann es für ein *Gespräch* im sozialen Stratum bedeutend sein, ob es sich um einen Streit, „Small Talk“ oder eine moderierte Diskussion handelt, wie emotional das Gespräch geführt wird oder ob alle Beteiligten dieselbe Sprache sprechen. Im biologischen Stratum hingegen liegt ein Fokus auf den Organen, die an der Lautproduktion und Lautwahrnehmung beteiligt sind, wohingegen im physikalischen Stratum Frequenz, Schallübertragung und Reaktionszeiten von Bedeutung sind.

Eine mittels einer Modellierungssprache erstellte Spezifikation zu einer Domäne wird in der Informatik auch als *Ontologie* bezeichnet, wobei Komplexität und Ausdrucksstärke einer *Ontologie* sehr unterschiedlich ausfallen können. Diese Ontologie ist die Trägermenge einer Wissensrepräsentation.

Bei der Modellierung einer Wissensrepräsentation mit dem Computer sollten nach Davis, Shrobe und Szolovits (1993) folgende fünf Prinzipien berücksichtigt werden:

24 „Wissen” kann hier sehr vereinfacht als „Daten im Kontext” aufgefasst werden. „Kontext” kann als semantisches Netzwerk aufgefasst werden (vgl. Sowa 2014).

25 Dieser Ausschnitt wird im Umfeld der Wissensrepräsentation Domäne genannt.

26 Wenn der Term „Sprache“ in diesem Kapitel verwendet wird, so ist damit, falls nicht anders erwähnt, eine formale Sprache gemeint.

27 Engl.: Knowledge Engineering.

28 In der Wissenschaft der Wissensrepräsentation wird für Gesichtspunkt häufig der Begriff Stratum verwendet (vgl. Poli 2001). Dabei sind soziologische, biologisch und physikalische Strata die am häufigsten verwendeten.

1. WISSENSREPRÄSENTATION IST EIN STELLVERTRETER:
   Dies Prinzip folgt den in *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism* (Ogden, Richards und Malinowski, 2013) vorgestellten semiotischen Verfahren. Der Computer repräsentiert echtweltliche Dinge oder Beziehungen durch Symbole und Relationen zwischen diesen Symbolen. Die Symbole, meistens in Form von Zeichenketten, sind die Stellvertreter der Entitäten einer Domäne und bilden die Grundlage des Modells. Da die Ausdrucksstärke von Computern beschränkt ist, sind diese Modelle nur Annäherungen oder Sichten auf die Domäne. Es ist Aufgabe des Domänenexperten und des Wissensdesigners (siehe Punkt 5 in dieser Aufzählung) zu entscheiden, wie umfassend eine Domäne zu modellieren ist und welche Symbole und Konzepte verwendet werden sollten.

2. WISSENSREPRÄSENTATION BESTEHT AUS EINER MENGE ONTOLOGISCHER FESTLEGUNGEN[29]:
   Diese Festlegungen bieten den philosophischen Überbau der zu modellierenden Domäne und legen damit das Vokabular, das für die Beschreibung einer Domäne in Frage kommt und dessen Bedeutung fest. So wird ein *Gespräch* im sozialen Stratum Kategorien wie *Sprache*, *Teilnehmer* oder *Gesprächsthema* benötigen, wohingegen im biologischen Stratum die beteiligen Organe klassifiziert werden sollten. Häufig geht das ONTOLOGICAL COMMITMENT über das Erstellen eines Vokabulars hinaus und beschäftigt sich auch mit der Existenz der Entitäten, die durch ein Vokabular beschrieben werden (vgl. Bricker 2016). Die verwendeten ontologischen Festlegungen sind abhängig von dem gängigen Wissensstand über die Domäne und einem philosophischen Überbau.

3. WISSENSREPRÄSENTATION SOLLTE SCHLUSSFOLGERUNGEN INNERHALB DER DOMÄNE ZULASSEN:
   In einem Wissensmodell sollten nicht nur die Dinge an sich, sondern auch deren Verhalten und Beziehungen abgebildet werden. Damit soll sich eine axiomatisierbare Theorie über eine Domäne bilden lassen, mit der Schlussfolgerungen möglich sind. Als axiomatisierbare Theorie ist dabei eine minimale Menge an Aussagen über eine Domäne zu betrachten, aus der sich alle weiteren Aussagen, die innerhalb dieser Theorie als „wahr" betrachtet werden können, ableiten lassen. Dies legt der zu modellierenden Wissensrepräsentation einige Einschränkungen auf. Damit wichtige Nebenbedingungen, wie die Widerspruchsfreiheit einer Theorie, erhalten bleiben, sind den bei der Modellierung verwendeten Sprachen gewisse Restriktionen auferlegt.

4. WISSENSREPRÄSENTATION SOLLTE FÜR EFFIZIENTE BERECHNUNGEN GEEIGNET SEIN:
   Anfragen an ein Wissensmodell sollten vom Computer in angemessener Zeit beantwortet werden. Zudem sollten diese Antworten innerhalb der Domäne korrekt und vollständig sein. Dieser Punkt ist ein

29 Engl.: Ontological Commitment.

nicht zu unterschätzendes Problem bei der Konzeptualisierung einer Domäne und hängt stark von der in Punkt 3 in dieser Aufzählung vorgestellten Axiomatisierung ab. Eine formale Sprache beschränkt nicht nur die Ausdrucksstärke, sondern setzt auch gewisse Grenzen bei der Berechnung eines komplexen Problems.

5. WISSENSREPRÄSENTATION IST EIN MITTEL MENSCHLICHER AUSDRUCKSWEISE:
   Wissen über eine Domäne ist an den aktuellen Forschungsstand und an das menschliche Vorstellungs- und Ausdrucksvermögen gebunden. Meistens sind Domänenexperte und Wissensdesigner nicht identisch. Es muss darauf geachtet werden, dass Wissensdesigner und Domänenexperte bei der Entwicklung einer Wissensrepräsentation einander verstehen. Dieser Prozess der Synchronisierung nimmt Zeit in Anspruch und ist für gewöhnlich iterativ.

## 1.2 EIN KURZER GESCHICHTLICHER ÜBERBLICK

Das Wort ONTOLOGIE leitet sich aus dem Altgriechchen *óntos* und *lógos* ab und bedeutet so viel wie *Wissenschaft des Seienden* und wurde Anfang des 17. Jahrhunderts erstmals von Rudolf Göckel und Jacob Lorhard in Schriften verwendet (vgl. Øhrstrøm, Andersen und Schärfe 2005). Als Teil der Metaphysik sind die dahinter stehenden Prinzipien natürlich viel älter. In der Philosophie ist ONTOLOGIE das Teilgebiet der Metaphysik, welches sich mit der Kategorisierung und Klassifizierung der „Dinge, die existieren“[30] beschäftigt (vgl. Inwagen und Sullivan 2017). Die erste Anwendung ontologischer Prinzipien wird Parmenides (um 500 v. Chr) nachgesagt. In seinem Werk *Über die Natur* führt er eine erste Unterscheidung von Dingen in „Seiende“ und „Nicht-Seiende“ (vgl. Parmenides 1986) ein. Die Grundlage für die Metaphysik an sich und ontologische Betrachtungen legte Aristoteles (384 v. Chr–322 v. Chr) in seinen Schriften *Kategorien* und *Metaphysik*. In *Kategorien* wird der erste ontologische Baum beschrieben (siehe Abbildung 1.1), der zwei Hauptkategorien umfasst; zum einen die *Substanz* und zum anderen den *Akzidens*. *Metaphysik* führt den Begriff des DING AN SICH (BEING QUA BEING) ein. Dieser Term gilt als ein zentraler Begriff der philosophischen Ontologie (vgl. Gracia 1999) und beschäftigt sich mit der Fragestellung, ob und welche Dinge unabhängig von anderen Dingen existieren können. Aristoteles hat auch in der Logik eine besondere Bedeutung. So gilt er als Begründer des Syllogismus und des Prinzips des logischen Widerspruchs (vgl. Aristoteles 1949). Diese beiden Prinzipien bilden die Grundlage für das logischen Schließen und sind damit auch in der computergestützten Wissensrepräsentation von Bedeutung.

30 „Ding“ ist eine etwas problematische Bezeichnung, da wir eine intuitive Vorstellung von dem Begriff „Ding“ als physikalisches Objekt haben, im Kontext der Ontologie sind aber auch nicht-physische Sachen, wie Vorstellungen oder Prozesse gemeint. Eine Alternative zu dem Begriff „Ding“ ist „Entität“. Die Verwendung im Deutschen beruht auf dem englischen Begriff „Thing”, welches in der formalen Sprache OWL als „Etwas, was in irgendeiner Form existiert“, definiert ist.

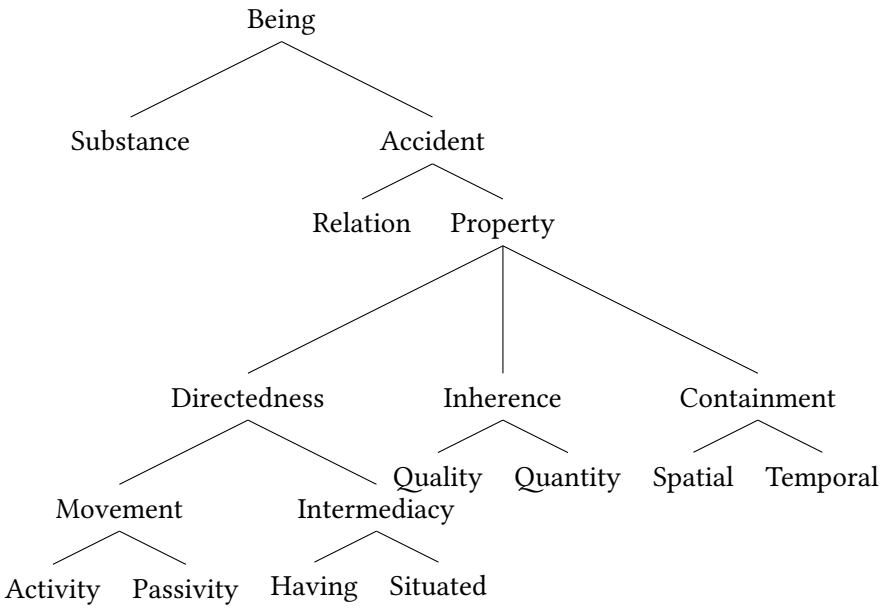



Abbildung 1.1: Die Hauptkategorien Aristoteles nach Sowa (2000).

Die Unterscheidung zwischen *Substanz* als Dinge, die unabhängig in der Welt existieren und *Akzidens* als Dinge, die nur in Abhängigkeit von anderen existieren, ist ein Motiv, das über die Jahrhunderte erhalten blieb und von mittelalterlichen Philosophen wie Thomas von Aquin (1225–1274) und Avicenna (980–1037) (vgl. Goodwin 1965; Janssens 2006) in ihren Werken zum Gottesbeweis aufgegriffen wurde.

Während der Aufklärung etablierte sich unter den Philosophen Christian Wolff (1679–1754) und Alexander Baumgarten (1714–1762) die Ontologie als philosophische Disziplin. Diese Epoche bot mit Wissenschaftlern wie Gottfried Wilhelm Leibniz (1646–1716) auch wichtige Erkenntnisse im Bereich der Mathematik, Logik und Graphentheorie. Als eine der ersten Domänenontologien in Form eines Graphen kann Carl von Linnés (1707–1778) Baum der Arten gesehen werden (vgl. Linné und Gmelin 1788).

Deutliche Kritik an der von Wolff und Baumgarten propagierten rationalen Sicht auf die Ontologie (vgl. Wolff 1963) als Wissenschaft übte Kant (1724–1804). Seine Schrift *Kritik der reinen Vernunft* (vgl. Mohr und Willaschek 2010) kritisiert den a-priori-Ansatz der bisherigen ontologischen Betrachtungen und postuliert, dass Kognition und Empfindung einen wichtigen Einfluss auf die Wahrnehmung der Welt haben, von der wir uns nicht einfach befreien können. Diese Ansicht wurde von den Philosophen des 19. Jahrhunderts aufgenommen. In der zweiten Hälfte des 19. Jahrhunderts wurde dann versucht, diese Ansichten in die aristotelische Auffassung von Ontologie zu integrieren.

Die Entwicklung der Boolschen Logik durch George Boole (1815–1864) und des Prädikatenkalküls durch Gottlob Frege (1848–1925) lieferten die theoretischen Grundsteine für Computer und die Informatik. Charles Sanders Peirce (1839–1914) entwickelte nicht nur den NAND-Operator, der eine

entscheidende Rolle in der Schaltkreisentwicklung spielt[31], und führte mit den Quantoren (∀, ∃) die Standardnotation der Prädikatenlogik ein, sondern leistete auch bedeutende Beiträge zur Semiotik, die für die Wissensrepräsentation im Computer unerlässlich ist. Seine Arbeiten zur Ontologie lieferten drei Basiskategorien, die sich auf die Arbeiten von Kant zur Ontologie zurückführen lassen. Die *Firstness* beschreibt konzeptuelle oder existenzunabhängige Dinge. *Secondness* beschreibt abhängige Dinge und *Thirdness* die Relationen zwischen Dingen der *Firstness* und der *Secondness* (vgl. Peirce 1974).

Das frühe 20. Jahrundert war geprägt von einer Formalisierung der philosophischen Betrachtungen. Husserls (1859–1938) Hauptwerk *Logische Untersuchungen* führt den Begriff der *Formalen Ontologie* ein, sein Zeitgenosse Stanisław Leśniewski (1886–1839) entwickelte mit der Mereologie eine Theorie zur Teil-Ganzes-Beziehung, was eine genauere Betrachtung von Beziehungen zwischen Dingen ermöglicht, die mit der klassischen Mengentheorie so nicht möglich ist (vgl. Leśniewski 1929).

In der zweiten Hälfte des 20. Jahrhunderts rückte der Computer als Werkzeug für die Wissensrepräsentation in den Fokus. Unter dem Begriff der „künstlichen Intelligenz" wurde ab den 60er-Jahren versucht, mithilfe von logischen Schlusssystemen „Intelligenz" im Computer zu erzeugen (vgl. McCarthy 1968; Minsky 1988). Obwohl diese Experimente hinter den Erwartungen zurück blieben, lieferten sie wichtige Ansätze im Bereich der Wissensrepräsentation. Insbesondere die Entwicklungen von Datenbanken und Metasprachen zur Wissenskodierung oder Beschreibung[32] können als Produkte dieser Zeit betrachtet werden. Auch stellte sich mit zunehmendem Einfluss des Computers in der Forschung heraus, dass für eine effiziente Anwendung des Computers das angesammelte Wissen besser strukturiert werden sollte und es Möglichkeiten bedarf, dieses Wissen über Institutionen hinweg verfügbar und zwischen Systemen kompatibel zu machen. Ontologien wurden als Möglichkeit gesehen, diese Anforderungen zu erfüllen, und sind besonders in Biologie und Medizin ein fester Bestandteil der Datenstrukturierung (vgl. Apweiler u. a. 2004; Stearns u. a. 2001). Mit dem Internet und dem damit verbundenen Vernetzen von Daten rückten graphenbasierte Datenstrukturen wieder in den Fokus. In *The Semantic Web* (vgl. Berners-Lee, Hendler und Lassila 2001) wird die Grundidee des „Semantischen Web" vorgestellt. Mit den von der W3C[33] entwickelten Sprachen RDF (1999) und dem bereits erwähnten OWL (2004) wurden Methoden entwickelt, Ontologien zu erstellen und Relationen zwischen Daten einheitlich und mit dem Internet als Zielplattform beschreiben zu können.

## 1.3 DER BEGRIFF *ONTOLOGIE*

Ontologie ist nicht nur ein fester Begriff in der Philosophie, sondern hat sich auch in der Informatik als Term etabliert (vgl. Horrocks 2013). Die Bedeutung in der Informatik weicht allerdings von der Bedeutung im philoso-

31 Der NAND-Operator ermöglicht es, Schaltkreise aus nur einem Grundbausstein aufzubauen.
32 Diese Metasprachen sind die bereits auf Seite 21 erwähnten *Beschreibungslogiken*.
33 World Wide Web Consortium, <https://www.w3.org>, abgerufen 05.02.2018.

phischen Kontext ab[34] und setzt einen deutlich stärkeren Schwerpunkt auf die Beschreibung einer Domäne mittels einer formalen Sprache. Häufig wird der Begriff Ontologie auch mit Klassifikationshierarchie oder einem kontrollierten Vokabular gleichgesetzt, was als Teilmengen oder Unterklassen des Begriffes Ontologie angesehen werden kann. Abbildung 1.2 zeigt eine Übersicht über das Spektrum des Ontologiebegriffs in der Informatik.

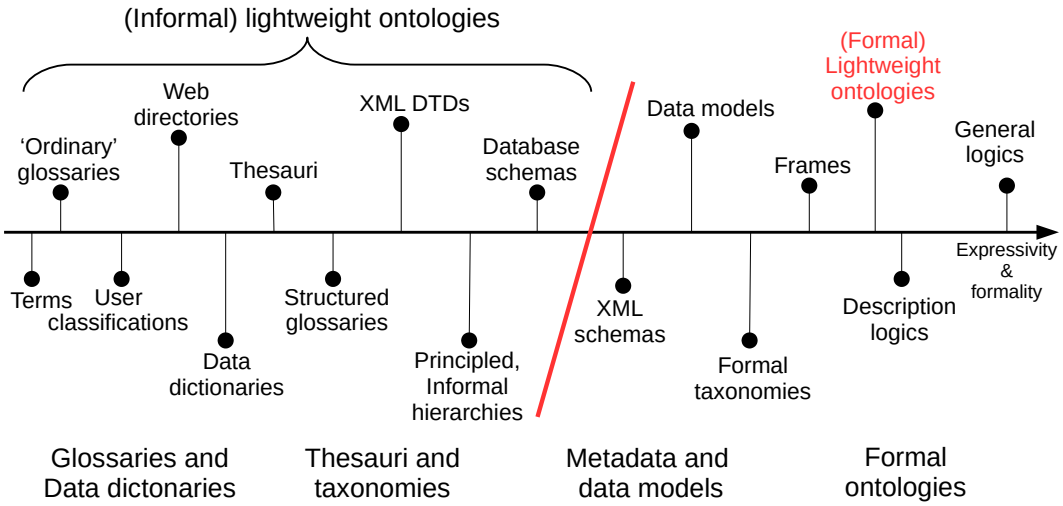



Abbildung 1.2: Übersicht über die verschiedenen Arten von Ontologien. Bild nach Giunchiglia und Zaihrayeu (2009), ursprüngliche Version von Uschold und Gruninger (2004).

In *Ontologies and Knowledge Bases: Towards a Terminological Clarification* (vgl. Guarino und Giaretta 1995) werden folgende sieben verschiedene Bedeutungen für den Begriff *Ontologie* aufgelistet:

1. Ontologie als philosophische Disziplin. Die Ontologie als „Wissenschaft des Seienden“ ist ein Teilgebiet der Metaphysik (vgl. Abschnitt 1.2).

2. Ontologie als ein informales System für Konzepte. Diese Definition entspricht unserer intuitiven Auffassung von Struktur und Hierarchie. Wir verwenden oft Klassifikationen, ohne uns zu tief mit der genauen Bedeutung auseinanderzusetzen. So sind wir in der Lage, Terme wie Katze, Säugetier, Mensch, Mann oder Frau in einen Kontext zu setzen, ohne uns tiefe Gedanken über die Bedeutung dieser Terme zu machen.

3. Ontologie als eine formale semantische Beschreibung. Diese Interpretation erweitert den vorherigen Punkt um eine Wissensgrundlage. Der Term Katze ist nun festgelegt und mit einer Beschreibung versehen.

4. Ontologie als eine Spezifikation einer Konzeptualisierung. Hier werden Begriffe nicht nur beschrieben, sondern klar und abgegrenzt definiert und in Beziehung zueinander gesetzt. Ein ontologischer Term ist nicht mehr nur über eine Beschreibung definiert, sondern auch über seine Beziehungen zu anderen Termen, wie in einem semantischen

34 Im Englischen wird diese Abgrenzung oft durch ***O**ntology* für die philosophischen und ***o**ntolgy* für die informatische Bedeutung kenntlich gemacht.

Netzwerk. Eine Konzeptualisierung kann zum Beispiel einschließen, dass Mann und Frau Menschen sind und dass Katzen und Menschen sich unterscheiden. Auch können den Begriffen Eigenschaften zugeordnet werden, zum Beispiel, dass Menschen aufrecht gehend sind, während Katzen auf vier Beinen gehen.

5. Ontologie als Repräsentant eines Systems für Konzepte mittels einer logischen Theorie
   a) beschrieben durch spezifische formale Eigenschaften,
   b) beschrieben durch einen spezifischen Zweck.

   Diese Interpretation beschreibt eine Ontologie als eine logische Theorie basierend auf einer formalen Sprache und einem dazugehörigen Axiomensystem. So kann zum Beispiel der Unterschied zwischen Katze und Mensch ausgedrückt werden als: $Mensch \cap Katze = \emptyset$ (Die Schnittmenge der Menge *Katze* mit der Menge *Mensch* ist disjunkt).

6. Ontologie als das Vokabular, das in einer logischen Theorie verwendet wird. Dies setzt den Begriff der Ontologie mit dem Begriff des Ontological Commitment (vgl. Seite 16) gleich.

7. Ontologie als die Metaspezifikation einer logischen Theorie. In dieser Interpretation stellt eine Ontologie die Grundbausteine für eine Theorie bereit.

Alle Punkte bis auf den ersten haben einen deutlichen Bezug zur Informatik und Wissensrepräsentation. Tatsächlich folgt Punkt 4 in der obigen Auflistung der am häufigsten zitierten Definition von Ontologie in der Informatik:

> „An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented." (Gruber 1995)

Ontologien mit einem hohen Grad an Formalisierung werden auch „formale Ontologien" genannt. Diese Ontologien bilden häufig eine Implementierung einer philosophischen Auffassung in einer formalen Sprache. Nino Cocchiarella definiert Formale Ontologie als:

> „Formal Ontology is a discipline in which the formal methods of mathematical logic are combined with the intuitive, philosophical analyses and principles of ontology." (Cocchiarella 2007)

Es ist nicht immer möglich eine formale Ontologie vollständig im Computer umzusetzen, da der Computer für gewöhnlich auf entscheidbare Teilmengen des Prädikatenkalküls beschränkt ist. Diese Teilmengen werden Beschreibungslogiken[35] genannt (vgl. Baader, Horrocks und Sattler 2008) und bilden seit den 60er-Jahren des 20. Jahrhunderts eine Gruppe von Sprachen, die explizit für die computergestützte Wissensrepräsentation entwickelt wurden. Seit 2004 bietet die von der W3C entwickelte OWL eine allgemeine Sprache

35 Engl.: Description logic.

zur Ontologieentwicklung. Formale Ontologien agieren oft als die Metaontologien für anwendungsbezogene Ontologien und liefern Konzepte und Kategorien für sehr grundlegende Entitäten, wie Prozess oder Raum und Zeit.

## 1.4 ONTOLOGIEN IN DER INFORMATIK

Ontologien in der Informatik haben im Gegensatz zur philosophischen Disziplin ein deutlich fokussierteres Ziel. Meistens sollen sie eine Strukturierung für eine Domäne liefern oder als Mediator zwischen verschiedenen Wissensrepräsentationen dienen. Man teilt Ontologien je nach Komplexität und Fokus in drei Gruppen ein:

TOP-LEVEL-ONTOLOGIEN versuchen die grundlegenden Dinge der Realität abzubilden. Sie liefern Formalisierungen für Raum und Zeit oder Teil-Ganzes-Beziehungen. Häufig versuchen sie eine metaphysische Grundannahme nachzubilden. Sie agieren als Meta-Ontologien für die folgenden und liefern die groben Rahmenbedingungen und Basiskategorien. Sie helfen dadurch auch beim Austausch von Informationen mithilfe von Ontologien[36]. Beispiele für Top-Level-Ontologien sind die GFO (General Formal Ontology) (vgl. Herre 2010; Herre u. a. 2007), DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (vgl. Gangemi u. a. 2002) oder SUMO (Suggested Upper-Merged Ontology) (vgl. Pease, Niles und Li 2002).

DOMÄNEN- UND AUFGABENONTOLOGIEN versuchen einen Teil der Welt (Domäne) so gut und vollständig wie möglich abzubilden. Idealerweise bedienen sie sich einer Top-Level-Ontologie und erweitern diese um die für die Domäne benötigten Konzepte. Je nach Definition der jeweiligen Domäne können diese Ontologien mehr oder weniger komplex oder genau ausfallen. Man spricht in diesem Zusammenhang auch von Granularität (vgl. Kumar, Smith und Novotny 2004). Für gewöhnlich werden diese Ontologien noch in *Upper-Domain*-Ontologies und Domänenontologien unterteilt, wobei die ersten einen eher allgemeinen Blick auf eine Domäne bieten und direkt mit den Top-Level-Ontologien interagieren (vgl. Beisswanger u. a. 2008), wohingegen die zweite deutlich spezifischer ist. Diese Arten von Ontologien finden häufig in der Wissenschaft Verwendung und dienen als Grundlage zur Datenstrukturierung oder Austausch. Eine Sammlung von Domänenontologien für die Medizin und Biologie findet sich im Bioportal[37] (vgl. Salvadores u. a. 2013).

ANWENDUNGSONTOLOGIEN erweitern oder beschränken die Domänenontologien, um sie in der Anwendung in Programmen oder Arbeitsabläufen nutzbar und praktikabel zu machen. Die meisten Ontologien, mit denen ein Endnutzer interagieren wird, fallen in diese Kategorie. In der Praxis sind Anwendungsontologien häufig kontrollierte Vokabulare oder Schemata, die eine hierarchische oder semantische Einord-

36 Top-Level-Ontologien sind der kleinste gemeinsame Nenner beim ONTOLOGICAL COMMITMENT.

37 <https://bioportal.bioontology.org>, abgerufen 05.02.2018.

nung zulassen. So kann der Dublin Core[38] als eine Anwendungsontologie gesehen werden, die einen Anwender darin unterstützt, Dokumente zu beschreiben. Die in dieser Arbeit vorgestellte *phonOntology* ist auch in erster Linie als eine Anwendungsontologie aufzufassen.

Abbildung 1.3 zeigt eine hierarchische Anordnung der verschiedenen Arten von Ontologien.

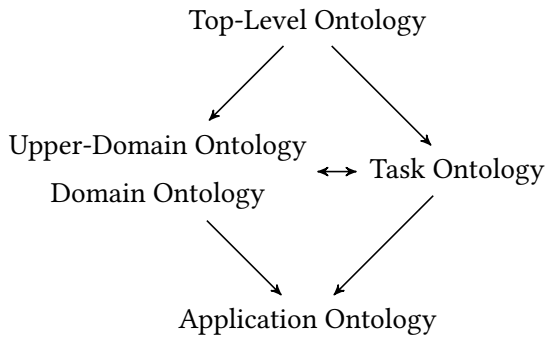



Abbildung 1.3: Verschiedene Arten von Ontologien nach Guarino (1997). Ein Pfeil kann als „Stellt Konzepte bereit"-Relation interpretiert werden.

In der Wissensrepräsentation hat sich ein spezielles Metavokabular für Ontologien entwickelt, um Terme einer Domäne beschreiben und klassifizieren zu können. Diese Beschreibungen genügen nicht notwendigerweise jeder philosophischen Ausrichtung oder entsprechen einem intuitiven Verständnis. Es ist außerdem schwierig, diese Terme zu definieren, da sie für gewöhnlich die Grundlage von Definitionen sind. Die Hauptterme sind:

ENTITÄT oder auch Ding symbolisiert die grundlegende ontologische Klasse. Alles Materielle als auch Immaterielle ist eine Entität. Die philosophischen Hintergründe sind auf Seite 17 kurz erwähnt. Eine *Drache* ist in diesem Zusammenhang eine Entität genauso wie eine tatsächlich existierende Tasse, die irgendwo in einem Regal steht.

KATEGORIEN sind abstrakte Entitäten, von denen sich andere Dinge instanziieren lassen. Mithilfe von Kategorien[39] ist es möglich, allgemeine Aussagen über eine Menge individueller Entitäten zu treffen und so zum Beispiel zu einem Oberbegriff zusammenzufassen oder mit anderen Kategorien in Relation zu setzen. Wie eine Kategorie gebildet wird, kann sich je nach zugrunde liegender philosophischer Auffassung unterscheiden. So kann eine Kategorie zum Beispiel als eine minimale Menge an Eigenschaften gesehen werden, die von einer Menge von Individuen inhäriert wird oder es können a-priori-Klassen sein, die unabhängig von Individuen existieren können. Beispiele für Kategorien sind *Mensch*, *Säugetier* oder *Stadt*. Eine Kategorie wird als Konzept bezeichnet, wenn wir dafür einen angemessenen linguistischen Ausdruck und eine direkte mentale Assoziation haben. So sind *Mensch* und *Stadt* auch Konzepte, aber ob es einen passenden linguistischen

38 <http://dublincore.org>, abgerufen 05.02.2018.

39 Im informatischen Kontext oft auch Klasse (engl. Class) genannt.

Ausdruck gibt, der die minimalen Eigenschaften einer zufälligen Menschenmenge beschreibt, ist nicht sicher.

INSTANZEN sind einzigartige, konkrete Entitäten, die keine Kategorien sind, aber durch Kategorien instanziiert werden. So kann das Individuum *Robert*[40] als eine Instanz der Kategorien *Mensch* oder *Säugetier* gesehen werden. Da Instanzen auch Entitäten sind, ist man nicht auf real existierende Dinge beschränkt. *Smaug*[41] kann also auch eine Instanz der Kategorie *Drache* sein.

RELATIONEN setzen Entitäten miteinander in Beziehung. Dadurch ist es möglich, Strukturen aufzubauen und Aussagen über diese Strukturen zu treffen. Welche Relationen benötigt werden, unterscheidet sich je nach gewählter Domäne, Stratum oder Granularität. Ein paar Relationen haben sich aber in der Wissensrepräsentation als de-facto-Standard durchgesetzt. Dazu gehören:

- Die Instanziierungsrelation, die verwendet wird um einem Individuum eine Kategorie zuzuordnen. So kann gesagt werden, dass *Robert* eine Instanz der Kategorie *Mensch* ist.
- Die Unterklassenrelation, mit der sich Kategorien genauer spezifizieren lassen. So kann die Kategorie *Mensch* als Unterklasse der Kategorie *Säugetier* gesehen werden. Häufig sagen wir „Ein Mensch *ist* ein Säugetier“ genauso wie „Robert *ist* ein Mensch“. Ontologisch haben diese beiden „*ist*“ aber zwei verschiedene Bedeutungen.
- Die Teil-Ganzes-Beziehung. Mereologische Relationen sind ein wichtiger Teil in einer ontologischen Strukturierung (vgl. Simons 2000). Diese Relationen erlauben Kompositionen von Entitäten zu komplexeren Entitäten oder beschreiben strukturelle Beziehungen. So kann man sagen: „Ein Dach *ist Teil* eines Hauses“ oder „Ein Wort *ist Teil* eines Satzes“. Die Teil-Ganzes-Relation wird meistens auf ein Stratum begrenzt, insbesondere wenn es um die Transitivität einer Relation geht. So kann man zwar sagen, diese Person *ist Teil* eines Gespräches und dieser Finger *ist Teil* dieser Person. Aber ob der Finger auch als Teil des Gespräches gesehen wird, ist Ansichtssache[42].

Eine besondere Eigenschaft einer formalen Axiomatisierung einer Ontologie ist die Möglichkeit, einen logischen Abschluss[43] zu bilden. Das bedeutet, auf Basis logischer Regeln lassen sich zusätzlich Informationen über Entitäten gewinnen. Ein Beispiel für Inferenz ist die Transitivität:

$$\forall x : x \in A \land A \subseteq B \rightarrow x \in B$$

40 Die Zeichenkette *Robert* ist in diesem Fall ein Repräsentant für eine real existierende Person.
41 Ein Drache aus dem Fantasieroman „Der Hobbit“.
42 Ontologisch kann sich Begriff „Person“ einmal auf einen Aktor im soziologischen Stratum oder auf die Repräsentation eines bestimmten Körpers im biologischen Stratum beziehen.
43 Auch Inferenz genannt.

Diese Formel besagt, dass für alle Elemente $x$ gilt, wenn $x$ ein Element der Menge $A$ ist und $A$ eine Teilmenge von $B$ ist. Dann folgt daraus, dass $x$ auch ein Element von $B$ ist.

Der logische Abschluss dient außerdem dazu, Ontologien auf logische Korrektheit zu überprüfen oder mögliche Widersprüche aufzudecken. Würden wir behaupten, dass *Robert* eine Instanz der Kategorien *Mensch* und *Katze* ist und an anderer Stelle, dass *Mensch* und *Katze* disjunkt sind, so würde Inferenz einen Widerspruch aufdecken. Im Idealfall ist eine Ontologie widerspruchsfrei.

Die Erstellung einer Ontologie erfolgt gewöhnlich in drei Schritten (vgl. Herre 2013), die iterativ durchlaufen werden. Diese Schritte leiten sich direkt aus den in Abschnitt 1.1 vorgestellten Punkten ab. Zuerst muss die Domäne abgegrenzt werden. Dabei müssen Entscheidungen zur Granularität und des Stratums getroffen werden. Der zweite Schritt umfasst das Ontological Commitment und die Konzeptualisierung (vgl. Guarino, Oberle und Staab 2009). Hier wird geklärt, welche Terme die Domäne benötigt und wie diese Terme aufzufassen sind. Im finalen Schritt wird die Ontologie axiomatisiert und in einer passenden Sprache implementiert. Die Axiomatisierung erfolgt auf den gewählten Termen und setzt diese Terme in eine logische Struktur, die sich möglichst gut in der gewählten Sprache abbilden lassen sollte.

## 1.5 DAS SEMANTISCHE WEB

Die Idee des Semantischen Webs basiert darauf, dass Internetadressen als Repräsentanten für Dinge agieren. Damit baut es auf der in Abschnitt 1.1, Punkt 1 vorgestellten Idee auf. Eine Internetadresse beschreibt damit nicht nur den Verweis auf eine Webseite, sondern kann auch als Repräsentant für ein *Ding* stehen. Solche Adressen werden auch Ressourcen[44] genannt. Links zwischen Ressourcen werden als Relationen aufgefasst. Zusätzlich können diese Ressourcen mit Metadaten versehen werden. Folgt diese Annotierung einem Standard, können automatisierte Agenten selbstständig diese Ressourcen entdecken und einordnen. Dadurch wird das Internet selbst zu einer Wissensbasis. In der Praxis bedeutet das, dass Webseiten zusätzliche Informationen innerhalb der HTML-Seite enthalten. Diese Informationen sind für gewöhnlich für den Nutzer versteckt, können aber von Agenten oder Webcrawlern gelesen werden. Eine weitere Möglichkeit ist, eine spezielle Version einer Ressource anzubieten. So kann ein menschlicher Benutzer eine HTML-Webseite angezeigt bekommen, während ein Agent ein spezielles maschinenlesbares Format lädt. Eines dieser Formate ist RDF.

### 1.5.1 *Resource Description Framework*

Die Resource Description Framework (RDF) ist ein Webstandard[45], der 1999 von dem World Wide Web Consortium (W3C) beschlossen wurde (vgl. Ma-

44 Engl.: Resource.

45 Da die W3C keine anerkanntes zwischenstaatliche Institution ist, gelten die Ausarbeitungen des W3C offiziell als Vorschläge (engl. Proposal.); es haben sich aber viele als „best practice“ Standard durchgesetzt.

nola und Miller 2004) und seit 2014 in der Version 1.1 vorliegt (vgl. Schreiber und Raimond 2014). RDF wurde als Strukturierungssprache für das Semantische Web entworfen und erlaubt, *Dinge* und Zusammenhänge (Relationen) in einfachen Aussagen der Form [<*subject*> <*predicate*> <*object*>] zu beschreiben. Diese Aussagen werden auch *Statements* oder *Triple* genannt. Eine Ansammlung von solchen Statements kann als Graph aufgefasst werden. Mathematisch ist ein Graph eine Menge von Ecken und Kanten (Vertices) :

$$G = (E, V) \text{ wobei gilt: } V \subseteq (E \times E)$$

Eine Kante ist also eine Relation[46] zwischen zwei Ecken. Anders ausgedrückt, eine Kante ist eine geordnete Menge von Ecken-Tupeln. In RDF modellierte Statements können als ein gerichteter Graph aufgefasst werden, wobei *Subjekt* und *Objekt* Elemente der Menge $E$ (Ecken) und *Prädikat* der Menge $V$ (Kanten) sind. Die Identifikation oder Benennung der Ecken und Kanten erfolgt über IRIs[47]. Die Auffassung einer Menge von RDF-Statements als mathematischer Graph erlaubt es, graphentheoretische Aussagen über diese Statements zu treffen und mathematische Operationen darauf anzuwenden. Eine wichtige Operation ist das Traversieren des Graphen. Damit kann überprüft werden, ob es einen Zusammenhang, genauer gesagt einen Kantenpfad, zwischen zwei Ecken gibt. So können komplexe Strukturen modelliert oder aufgedeckt werden. Auch erlaubt das Traversieren eines Graphen das Beantworten von Fragen an den Graph. Dadurch können diese Strukturen als Grundlage für Datenbanken verwendete werden (siehe Abschnitt 1.6).

Ein gültiges Statement in RDF unterliegt folgenden Annahmen:

- An jeder Position (<*subject*> <*predicate*> <*object*>) kann eine IRI vorkommen.
- Ein <*object*> kann auch ein Literal sein.
- Ein <*subject*> oder <*object*> kann eine Blank Node sein.

Eine IRI (vgl. Duerst und Suignard 2005) ist ein standardisiertes Format für dereferenzierbare Identifikatoren. Es stellt eine Generalisierung der URI (Uniform Resource Identifier) (vgl. Berners-Lee, Fielding und Masinter 1998) dar, die in Form von Internetadressen (URLs) bekannt sind. Der Hauptunterschied zwischen einer IRI und einer URI ist, dass eine IRI auch andere Schreibsysteme (Unicode) zulässt, wohingegen eine URI auf die Zeichen der amerikanischen Tastatur (ASCII[48]) beschränkt ist.

Eine URI[49] für RDF besteht aus:

```
<Protokoll>://<Domäne>/<Kontext></ oder #><Identifikator>
```

Ein `Protokoll` ist zum Beispiel das *Hypertext Transfer Protocol* (vgl. Fielding u. a. 1999), abgekürzt mit *http*; eine `Domäne` der Name einer Website, wie zum Beispiel *issg.de* oder *www.regionalsprache.de*, ein `Kontext`[50] ist eine durch / getrennte genauere Spezifikation der Ressource und der `Identifikator` ist der letzte Teil der URI, abgegrenzt entweder durch ein / oder eine #. Bei einer Serialisierung von RDF können <Protokoll><Domäne>/<Kontext>

46 Mengentheoretisch wird eine Relation als geordnete Teilmenge des kartesischen Produkts von zwei oder mehr Mengen gesehen.

47 Internationalized Resource Identifier.

48 American Standard Code for Information Interchange.

49 Durch die Verbreitung von URI als Bezeichner von Internetadressen und der Nähe von RDF zum Semantischen Web werden in der Praxis meistens URI anstelle von IRI verwendet.

50 Häufig auch als Pfad bezeichnet.

durch ein Präfix abgekürzt werden. Dadurch werden sogenannte Namensräume definiert. Diese dienen in erster Linie der Lesbarkeit und zur Platzeinsparung bei der Datenübertragung. So kann beispielsweise die URI <http://issg.de/ontologies/phonetic#Vowel>, die sich aufteilt in das Protokoll *http*, die Domäne *issg.de*, den Kontext *ontologies/phonetic*, das Trennsymbol # und den Identifikator *Vowel*, verkürzt als *phon:Vowel* geschrieben werden, wobei *phon* als Präfix definiert wird und damit den Namensraum, in dem *Vowel* eine Ressource ist, bildet. Im Laufe der Zeit haben sich bestimmte Namensräume als Standard durchgesetzt, so dass sich damit semantische Bedeutungen assoziieren lassen. Tabelle 1.1 zeigt häufig verwendete Namensräume und ihre Abkürzungen. *Rdf*, *rdfs*, *owl* und *xml* sind kontrollierte Vokabulare, die zur besseren Strukturierung und Formulierung von RDF dienen, *dcterm* und *skos* sind Ontologien, die zur Beschreibung von Metadaten oder dem Wissensmanagement im Generellen dienen.

Tabelle 1.1: Häufig verwendete Namensräume und zugehörige Präfixe für RDF.

| PRÄFIX | URI |
|---|---|
| rdf: | <http://www.w3.org/1999/02/22-rdf-syntax-ns#> |
| rdfs: | <http://www.w3.org/2000/01/rdf-schema#> |
| owl: | <http://www.w3.org/2002/07/owl#> |
| xml: | <http://www.w3.org/XML/1998/namespace> |
| dcterm: | <http://purl.org/dc/terms/> |
| skos: | <http://www.w3.org/2004/02/skos/core#> |

Die Verwendung von URIs als Identifikatoren in RDF schlägt eine Brücke zum Internet und erlaubt es, Webadressen als Identifikatoren in Statements zu verwenden. Dadurch werden diese Adressen Teil der zu modellierenden Datenstruktur und können damit als Repräsentant einer realweltlichen Entität angesehen werden. Dies ist eine der wesentlichen Ideen hinter dem Semantischen Web.

Eine Blank Node ist eine nicht näher bezeichnete Ressource oder anonyme Ressource, die nur durch ihre Relationen beschrieben wird. Beim Erstellen von RDF-Graphen sollten Blank Nodes vermieden werden. Sie können aber als Ergebnis von Transformationen auf dem Graph, wie zum Beispiel Inferenz, auftreten oder als Repräsentant komplexer Entitäten dienen, für die eine einfache Benennung durch eine IRI nicht möglich oder sinnvoll ist.

Ein Literal ist alles, was keine IRI oder Blank Node ist und kann als der inhaltliche Datenträger betrachtet werden. Literale können Texte oder Zahlen, aber auch Binärdaten sein. Einem Literal kann ein Datenformat oder eine Sprache zugeordnet werden.

Eine Menge von RDF-Statements lässt sich in verschiedenen Formaten serialisieren. Manche Formate wie XML (vgl. Schreiber und Gandon 2014) oder JSON-LD (vgl. Lanthaler, Sporny und Kellogg 2014) sind für Menschen schwer zu lesen, andere wie TURTLE (vgl. Prud'hommeaux und Carothers

2014) haben eine durchaus verständliche Syntax. Nach einer Definition von Präfixen am Anfang der Datei, werden RDF-Statements in der [*<subject> <predicate> <object>*] Form in eine Zeile geschrieben und mit einem Punkt abgeschlossen. Kommentare werden durch eine führende # gekennzeichnet und dienen ausschließlich als Annotation für einen menschlichen Leser. Der resultierende Graph aus der folgenden Serialisierung in TURTLE ist in Abbildung 1.4 gezeigt:

```
PREFIX urn: <urn:example.net/> # Namensraumdefinition
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
urn:id27362 rdf:type urn:Mensch .
# urn:id27362 ist eine willkürlich generierte ID, die die
↪  Person Robert repräsentieren soll.
urn:id27362 rdfs:label 'Robert' .
# 'Robert' ist ein Literal
urn:Mensch rdfs:subClassOf urn:Mammal .
urn:Marburg rdf:type urn:City .
```

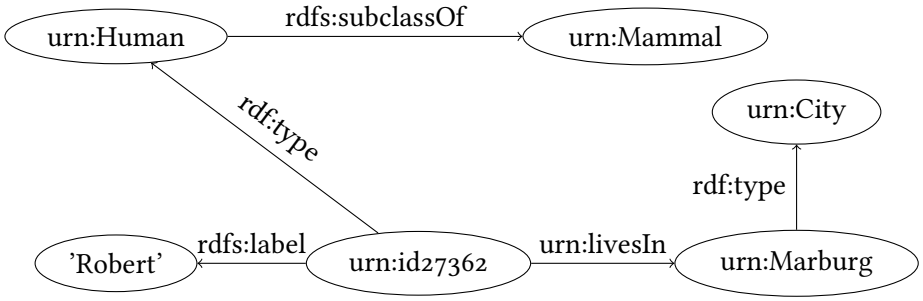



Abbildung 1.4: Die TURTLE-Serialisierung als Graph.

Dieser Graph beschreibt eine Ressource markiert durch die URI *<urn:id27362>*. Dieser Ressource wird mittels der Relation *rdfs:label* das Literal „Robert“ in Form eines Textstrings zugeordnet. Außerdem wird über die Relation *rdf:type* ausgedrückt, dass sie eine Instanz der Kategorie *urn:Human* ist und über die Relation *urn:livesIn* wird eine Beziehung zu der Ressource *urn:Marburg* hergestellt, die wiederum eine Instanz der Kategorie *urn:City* ist. Schließlich wird noch festgelegt, dass *urn:Human* eine Unterklasse von *urn:Mammal* ist.

Eine Erweiterung zu RDF-Graphen sind *Datasets*[51]. Um viele RDF-Statements besser kontrollieren zu können, werden die RDF-Triple in einen Kontext in Form eines *Named Graphs* gesetzt. Dadurch werden aus den Triple Quads, beziehungsweise aus *Subjekt Prädikat Objekt* wird *Subjekt Prädikat Objekt Kontext*. So lassen sich Statements besser gruppieren. Ein RDF-Datenset besteht aus o bis n *Named Graph* und einem *Default Graph*. Diesem Graph ist kein Kontext zugeordnet und wird bei Anfragen gegen das RDF-Datenset standardmäßig verwendet. Der *Default Graph* ist in den meisten Fällen die Vereinigung aller *Named Graphs*. RDF-Datensets sind die Im-

51 Da in Kapitel 3 und folgenden der Begriff „Dataset“ auch eine wichtige Rolle spielt, allerdings in einem anderen Kontext, wird *Datenset* im weiteren als RDF-Datenset bezeichnet.

plementierungsgrundlage für eine besondere Form von Graphdatenbanken, den sogenannten TripleStores, auf die in Abschnitt 1.6 genauer eingegangen wird.

### 1.5.2 *Resource Description Framework Schema*

RDFS (Resource Description Framework Schema) (vgl. Guha und Brickley 2014) ist eine Erweiterung zu RDF. Sie dient in erster Linie dazu, Schemata für RDF-Daten zu entwickeln. Dazu stellt RDFS ein kontrolliertes Vokabular zur Verfügung, um zum einen eine Datenstruktur aufbauen zu können und zum anderen normierte Relationen für die Annotation von RDF bereitzustellen. RDFS erweitert RDF um eine leichte Semantik und führt ontologische Konzepte wie Klassen (Kategorien) und die Instanziierung ein. Die in RDFS eingeführten Relationen und Klassen werden unter den *rdf* und *rdfs* Namensräumen geführt. Die wichtigsten Klassen und Relationen sind:

- Alle Terme in RDF sind Instanzen der Klasse *rdfs:Resource*. Damit fungiert sie als Basisklasse von der sich alles ableitet. *rdfs:Resource* selbst ist vom Typ *rdfs:Class*.
- *rdfs:Class* ist die Basisklasse für alle Konzepte oder Kategorien in einem Schema. Elemente von *rdfs:Class* können instanziiert werden.
- *rdf:Property* ist die Hauptklasse für alle Prädikate oder Relationen. Sie ist auch eine Instanz von *rdfs:Class*.
- *rdf:type* ist die Instanziierungsrelation. Ein Statement der Form

  *A rdf:type B .*

  impliziert, dass *A* eine *rdfs:Class* ist. In vielen RDF-Formaten wird rdf:type für gewöhnlich mit einem einfachen „a“ abgekürzt. Zum Beispiel: *ex:Marburg a ex:City .*
- *rdfs:subClassOf* ist die transitive Unterklassenrelation und kann genutzt werden, um Konzepte in einer Baumstruktur anzuordnen. Eine Aussage der Form *A rdfs:subClassOf B* bedeutet, dass alle A und B *rdfs:Class* sind und dass alle Instanzen von A auch Instanzen von B sind.
- *rdfs:subPropertyOf* ist die analoge Relation zu *rdfs:subClassOf* für *rdfs:Property*. Diese Relation dient dazu, selbst definierte *Properties* in einer Baumstruktur anzuordnen.
- *rdfs:label* ist eine Annotationsrelation und wird verwendet um Ressourcen mit einem für Menschen verständlichen Bezeichner (Label) zu versehen. Da in einem RDF-Datenset alle Entitäten einen eindeutigen Identifikator (IRI) benötigen, ist es nicht immer möglich oder sinnvoll, diesen Identifikator für den Menschen verständlich zu wählen. *Rdfs:label* bietet eine normierte Möglichkeit, verständliche Bezeichner einer Ressource hinzuzufügen.
  So steht die URI <https://www.uniprot.org/taxonomy/9606>[52] für die

52 Abgerufen 05.03.2018.

taxonomische Klasse „Homo Sapiens“ in der Uniprot Datenbank[53]. Als RDF-Fragment kann dies als

```
<https://www.uniprot.org/taxonomy/9606>
            rdfs:label "Homo Sapiens".
```

ausgedrückt werden. Label müssen nicht einzigartig oder eindeutig sein. Häufig gibt es mehrere Label zu einer Ressource. Insbesondere wenn Daten für mehrere Sprachen annotiert sind.

- *rdfs:range*, *rdfs:domain* sind zwei Relationen, mit denen eine Relation eingeschränkt werden kann. Dass die Entitäten, die durch sie verbunden werden, bestimmten Klassen angehören

```
urn:livesIn rdfs:range urn:City .
urn:livesIn rdfs:domain urn:Human .
```

bedeutet, dass alle Subjekte A in *A urn:livesIn B* Instanzen der Klasse *urn:Human* sind und alle Objekte B Instanzen der Klasse*urn:City*.

Mit *rdfs:subClassOf*, *rdfs:subPropertyOf*, *rdfs:domain* und *rdfs:range* bietet RDFS Inferenzmöglichkeiten. So lassen sich mit *rdfs:domain* und *rdfs:range* Instanzen implizit einer Klasse zuordnen. Die beiden *sub*-Relationen ermöglichen eine Hierarchisierung von Klassen und eine implizite Weitergabe dieser Struktur an die Instanzen einer Subklasse.

RDFS folgt der Offene-Welt-Annahme[54]. Das bedeutet, dass solange eine Aussage nicht explizit als Widerspruch markiert wird, ist diese gültig. Eine Aussagenmenge wie:

```
urn:livesIn rdfs:range urn:City .
urn:livesIn rdfs:domain urn:Human .
urn:id27362 rdf:type urn:Human .
urn:Marburg rdf:type urn:city .
urn:Marburg urn:livesIn urn:id27362 .
```

würde implizieren, dass *urn:Marburg* zusätzlich zur Stadt auch ein Mensch ist und *urn:id27362* sowohl Mensch als auch Stadt ist. Es ist in RDFS nicht möglich, solche Konstrukte in der Sprache selbst zu verbieten. Erst eine komplexere Logik wie OWL[55] ermöglicht es, Restriktionen zu definieren. Das bedeutet auch, dass eine Aussage wie oben nicht als semantischer Widerspruch erkannt wird, und bei einem unachtsamen Design kann dies Auswirkungen auf die gesamte Struktur der Daten haben.

### 1.5.3 *Beschreibungslogik*

Mit RDFS ist es möglich, RDF-Aussagen eine einfache Semantik zu geben. Komplexere Aussagen erfordern allerdings ausdrucksstärkere Sprachen. Eine dieser Sprachen ist die Web Ontology Language. OWL gehört zu den Beschreibungslogiken (vgl. Baader, Horrocks und Sattler 2005). Eine Beschreibungslogik ist ein Fragment des Prädikatenkalküls der ersten Stufe. Dabei

53 <https://www.uniprot.org> Abgerufen 05.03.2018.
54 Engl.: Open-World-Assumption (OWA).
55 Web Ontology Language (kurz: OWL).

wird besonderer Wert auf die Entscheidbarkeit[56] der dadurch produzierbaren Ausdrücke gelegt. Beschreibungslogiken wurden explizit für die Wissensrepräsentation entwickelt. Ein Fokus ist dabei neben der Entscheidbarkeit auch die Möglichkeit zur Schlussfolgerung. Es gibt verschiedene Ausprägungen von Beschreibungslogiken, die unterschiedliche semantische Einschränkungen haben können, aber je nach Anwendungsgebiet für die Modellierung einer Domäne besser geeignet sein können.

### 1.5.4 *Web Ontology Language*

Die Web Ontology Language (vgl. Hitzler u. a. 2009; W3C OWL Working Group 2012) ist eine Erweiterung zu RDFS und ermöglicht als Beschreibungslogik die Entwicklung komplexer Ontologien. Die erste Version wurde 2004 vorgestellt und seit 2012 liegt OWL in der Version 2 als OWL2 vor, die die Ausdrucksmöglichkeiten noch einmal erweitert. Aussagen in OWL sind nicht notwendigerweise entscheidbar. Um dennoch effiziente Inferenz und computergestütztes Ontologiedesign mit OWL2 zu ermöglichen, wurden sogenannte Profile eingeführt, die gewissen Beschränkungen in der Ausdrucksstärke unterliegen, dafür aber Entscheidbarkeit in akzeptabler Zeit gewährleisten.

Es werden zudem neue Klassen und Relationen eingeführt. Alle Klassen und Relationen von RDFS sind auch Teil von OWL2, aber während RDFS hauptsächlich zur einfachen Datenstrukturierung gedacht ist, sind in OWL2 auch ontologische (und logische) Aspekte berücksichtigt Die wichtigsten Klassen und Relationen in OWL2 sind:

- *owl:Thing* ist die universelle Klasse. Jede benutzerdefinierte Klasse ist eine Unterklasse von *owl:Thing*. Daraus folgt direkt, dass alle Individuen Instanzen der Klasse *owl:Thing* sind. Ontologisch kann *owl:Thing* als Entität aufgefasst werden. *Owl:Thing* ist eine Instanz von *owl:Class*, die äquivalent zu *rdfs:Class* ist.

- *owl:Nothing* ist die „leere“ Klasse und damit das semantische Gegenstück zu *owl:Thing*. *Owl:Nothing* ist eine Unterklasse jeder Klasse in OWL und kein Individuum darf Instanz dieser Klasse sein. Diese Klasse sollte nicht explizit verwendet werden, sondern höchstens nach der Anwendung von Inferenz auftauchen, wenn dadurch ein logischer Widerspruch (Inkonsistenz) in der Ontologie aufgedeckt wurde. Eine Ontologie, in der Instanzen der Klasse *owl:Nothing* existieren, beinhaltet einen logischen Widerspruch und sollte überprüft werden. Das kann bedeuten, dass dieser Widerspruch entweder bei dem Design der Ontologie auftritt oder dass Daten widersprüchlichen Klassen zugeordnet wurden.

- *owl:disjointWith* ermöglicht zwei Klassen explizit distinkt zu machen und löst damit das in Unterabschnitt 1.5.2 geschilderte Problem, dass

56 Entscheidbarkeit bedeutet vereinfacht ausgedrückt, dass es einen Algorithmus gibt, der in endlicher Zeit berechnen kann, ob ein Ausdruck bei einer Variablenbelegung wahr oder falsch ist. Entscheidbarkeit wird für gewöhnlich mithilfe von Turing-Maschinen (vgl. Turing 1937) definiert und diese sind ein zentraler Gegenstand der theoretischen Informatik.

die Offene-Welt-Annahme bei RDFS *ex:Marburg* sowohl als Stadt als auch als Mensch klassifizieren kann. Wenn *ex:City* und *ex:Human* als disjunkt erklärt werden, führt die obige Aussage zu einem Widerspruch, genauer gesagt *ex:Marburg* wird als eine Instanz der Klasse *owl:Nothing* inferiert und dies widerspricht der Definition von *owl:Nothing*.

- *owl:equivalentClass* definiert zwei Klassen als gleich. Es ist zu beachten, dass in OWL2 eine Klasse auch eine komplexe Klasse sein kann, die sich aus mehreren logischen Konstrukten zusammensetzt. Dies macht diese Relation sehr mächtig, weil es das „Zerlegen" einer benannten Klasse in ihre Einzelteile ermöglicht. Diese Relation ist für die *phonOntology* von entscheidender Bedeutung, da durch sie ein Laut als eine Menge von Lauteigenschaften definiert werden kann.

- *owl:TransitiveProperty* definiert eine Relation als transitiv. Dies ist in RDFS der *rdfs:subClassOf* Relation vorbehalten. In OWL2 können nun auch andere Relationen als transitiv definiert werden, was viele neue Anwendungsmöglichkeiten eröffnet. Insbesondere lässt sich nun auch die mereolgische *partOf* -Relation als transitiv definieren.

- *owl:inverseOf* definiert zwei Relationen invers zueinander. So kann zum Beispiel „*ex:partOf owl:inverseOf ex:hasPart* ." definiert werden. Damit reicht es beim Erstellen der Statements aus, nur eine Relation explizit anzugeben. Die andere wird mittels Inferenz automatisch hinzugefügt.

- *owl:Ontology* deklariert eine Datei als Ontologie. Dadurch lassen sich Ontologien und Daten voneinander trennen. In der Praxis macht es Sinn, eine Ontologie zusätzlich noch in einem *Named Graph* zu speichern. In einem graphenbasierten Wissenssystem ist es nicht notwendig diese Unterscheidung vorzunehmen, aber sie ist sehr hilfreich bei der Verwaltung eines Wissenssystems.

Aus Anwendungsfällen bei der Benutzung von Ontologien haben sich drei Profile ergeben, um bestimmte Anwendungsfälle besser lösen zu können. Die Profile sind OWL2 EL, OWL2 QL und OWL2 RL, wobei jedes Profil einen bestimmten Fokus setzt und damit gewisse Restriktionen an die Ontologieentwicklung stellt.

OWL2 EL ist hilfreich bei der Entwicklung von umfangreichen, baumstrukturartigen Ontologien, wie sie häufig in der Biologie oder Medizin vorkommen. Eine wichtige Einschränkung ist, dass die *owl:inverseOf* Relation nicht verwendet werden darf. Damit muss die Richtung immer explizit angegeben werden, allerdings können *rdfs:range* und *rdfs:domain* inferiert werden. OWL2 EL basiert auf der $\mathcal{EL}$ Beschreibungslogik (vgl. Baader, Brandt und Lutz 2005).

OWL2 QL ist ein Profil, das sich an den traditionellen relationalen Datenbanken orientiert. QL steht für *Query Language* und soll auf Anfragesprachen wie SQL hinweisen. Dieses Profil verwendet weitestgehend die Relationen und Klassen von RDFS, ignoriert allerdings Transitivität. Dieses Profil ist hilfreich, wenn man Kompatibilität zu relationalen Datenbanken wahren muss.

Das OWL2 RL Profil ist dafür gedacht, auf Basis von Regeln das Wissen über die Daten zu erweitern. Es bietet damit die größten Inferenzmöglichkeiten. Einschränkungen gibt es aber bei der Reflexivität[57] und bei der Definition fixer Mengenangaben. RL steht für *Rule Language* und bietet ein regelbasiertes Axiomsystem auf Basis von Implikationen[58] in Form von Horn-Klauseln (vgl. Horrocks u. a. 2005) zur Wissenserweiterung an. Das Profil ist sprachneutral, das heißt, dass die Regeln für OWL2 RL in verschiedenen formalen Sprachen verfasst werden können. Meistens bieten sich dafür logikbasierte Programmiersprachen wie Prolog (vgl. Clocksin und Mellish 2012) oder SWRL (vgl. Horrocks u. a. 2004) an. Über diese Regeln können alle erlaubten Klassen und Relation in OWL2 RL definiert werden. In Tabelle 1.2 werden einige ausgewählte Regeln vorgestellt. Eine vollständige Definition findet sich in Motik u. a. (2012).

Tabelle 1.2: Ausgewählte OWL2 RL Regeln. Dabei sind die Inhalte von WENN … und DANN … als RDF-Triple zu lesen (x, y, p, z, c1, c2 sind Platzhalter für beliebige URIs).

| REGEL | WENN … | DANN … |
|---|---|---|
| Symmetrie | x owl:sameAs y . | y owl:sameAs x . |
| Transitivität | p a owl:TransitiveProperty .<br>x p y .<br>y p z . | x p z . |
| Klassenäquivalenz 1 | c1 rdfs:subClassOf c2 .<br>c2 rdfs:subClassOf c1 . | c1 owl:equivalentClass c2 . |
| Klassenäquivalenz 2 | c1 owl:equivalentClass c2 .<br>x a c1 . | x a c2 . |
| Leere Klasse | x a owl:Nothing . | WIDERSPRUCH |
| Klassendisjunktheit | c1 owl:disjointWith c2 .<br>x a c1 .<br>x a c2 . | WIDERSPRUCH |

Mithilfe der in OWL2 definierten Semantik lassen sich Klassen, Relationen und Axiome definieren. All dies zusammen ist eine Ontologie in der Informatik. Diese Ontologie kann auf ein RDF-Datenset angewendet werden, um das RDF-Datenset mit einer Struktur zu versehen oder durch inferiertes Wissen zu erweitern. Im Kontext der Beschreibungslogik wird die Ontologie auch TBox (Terminological Box) und das Datenset ABox (Assertional Box) genannt.

## 1.6 GRAPHDATENBANKEN

Zu einem Wissenssystem gehören neben einem Schema (oder Ontologie) auch Daten und ein System, diese Daten effizient zu verwalten und verfügbar zu machen. Solche Systeme werden unter dem Begriff Datenbank oder

57 Man kann in OWL2 RL nicht definieren, dass ein Ding mit sich selbst in Beziehung steht. Zum Beispiel kann man nicht explizit ausdrücken, dass jede Person sich selbst kennt.

58 „Wenn-dann"-Regeln.

Datenbankmanagementsystem geführt. Der bekannteste Vertreter von Datenbanken ist die relationale Datenbank mit der Anfragesprache SQL. Relationale Datenbanken sind sehr verbreitet und dominieren den Datenbankmarkt. In jüngerer Zeit haben andere Datenbankmodelle unter dem Begriff der NoSQL[59]-Datenbanken vermehrt Zulauf bekommen. Bekannte Beispiele für NoSQL-Datenbanken sind Objekt-Datenbanken wie MongoDB[60] oder Redis[61] oder Graphdatenbanken wie Neo4j[62] oder GraphDB[63]. Eine besondere Unterklasse von Graphdatenbanken sind die TripleStores. Graphdatenbanken sind zunächst schemalos. Eine Datenstruktur ergibt sich nicht wie bei relationalen Datenbanken aus einem vorher festgelegten Schema, sonder implizit aus den tatsächlich vorhanden Daten. Dies macht sie zu einer guten Wahl für heterogene Datensätze. Anders ausgedrückt bedeutet dies, dass es keine explizite Trennung zwischen einem Schemabereich und einem Datenbereich gibt. Ein Schema ist damit Teil der Daten selbst. Da Ontologien in RDF implementiert werden und RDF eine Graphenstruktur ist, bieten sich Graphdatenbanken an, um RDF-Daten und zugehörige Ontologien zu speichern. Dies bietet ein hohes Maß an Flexibilität für die Daten und die Ontologien können komplexe Schemata zu diesen Daten liefern. Im Folgenden ist ein Beispiel für Daten und Schema als RDF-Datensatz angeführt:

```
# SCHEMA
urn:PopulatedPlace rdfs:subClassOf geo:Feature .
urn:City rdfs:subClassOf urn:PopulatedPlace .
urn:LingObs rdfs:subClassOf qb:Observation .
urn:at rdfs:range geo:Feature  .
urn:at rdfs:domain qb:Observation .
# DATEN
urn:id234 urn:at urn:Marburg .
urn:id234 rdfs:label "e" .
urn:id234 a urn:LingObs .
urn:Marburg a urn:City .
urn:id235 urn:at urn:Marburg .
urn:id235 rdfs:label "25°C" .
```

Dieser Datensatz definiert eine Schemastruktur, die eine Stadt als Unterklasse von bewohnten Orten definiert, die wiederum eine Unterklasse von Landmarken ist. Linguistische Observation ist eine Unterklasse einer allgemeinen Observation. Durch die Einschränkungen der Relation *urn:at* wird ausgedrückt, dass Observationen an Landmarken vorkommen können. In den Daten werden Ressourcen nun verknüpft. Die Ressource *urn:id234* hat den Bezeichner „e“ und ist mit der Ressource *urn:Marburg* über die Relation *urn:at* verknüpft. Zudem wird festgelegt, dass *urn:id234* eine linguistische Observation ist und *urn:Marburg* eine Stadt. Die Ressource *urn:id235* hat keinen expliziten Typ, man kann aber inferieren, dass sie mindestens eine Observation ist, da sie mit einer Stadt über die Relation *urn:at* verknüpft ist.

59 NoSQL steht für *Not only SQL*.
60 <https://www.mongodb.com>, abgerufen 07.02.2018.
61 <https://redis.io>, abgerufen 07.02.2018.
62 <https://neo4j.com>, abgerufen 07.02.2018.
63 <http://graphdb.ontotext.com>, abgerufen 07.02.2018.

So dienen Ontologien dazu, die Daten durch die in der Ontologie definierten Regeln zu erweitern und in eine Struktur zu bringen, aber gleichzeitig kann das implizite Schema der RDF-Daten erhalten bleiben. So kann in der Ontologie definiert werden, dass jede Sprachkarte auch eine Karte ist. Ein optionaler Kommentar zu einer Karte benötigt aber keine genauere Klassifikation und muss nicht gesondert in der Ontologie beschrieben werden, sondern bildet einen Teil des impliziten Schemas auf den RDF-Daten. Umgekehrt bietet es auch die Möglichkeit, ein Schema aus heterogenen Daten zu extrahieren und anschließend in einer Ontologie zu strukturieren. So können Datensätze Bezeichnungsinformationen in einem *:name-* oder *:title*-Feld enthalten. In der Ontologie kann dann definiert werden, dass diese beiden Relationen Unterklassen von *rdfs:label* sind. Dies kann helfen, Anfragen an die Datenbank zu vereinheitlichen und unterschiedliche Datensätze untereinander kompatibel zu machen.

TripleStores sind Graphdatenbanken, die speziell darauf ausgelegt wurden, RDF-Daten zu verwalten. Mithilfe der Anfragesprache SPARQL (vgl. Seaborne und Harris 2013) kann auf diese Datenbanken zugegriffen werden. SPARQL wurde als eine Anfragesprache für das Internet entwickelt und ähnelt sehr der bereits vorgestellten Syntax von TURTLE. Eine einfache SPARQL-Anfrage ist im Folgenden gezeigt:

```
PREFIX urn: <urn:example.net/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select ?obs ?class  ?label where
{
?obs urn:at urn:Marburg .
?obs a ?class .
?obs rdfs:label ?label
}
```

Diese Anfrage sucht[64] nach allen „Dingen“, die über die Relation *urn:at* mit der Ressource *urn:Marburg* verbunden sind, eine Klasse haben und mittels *rdfs:label* mit einer weiteren Ressource verbunden sind. Das Ergebnis einer solchen Anfrage ist eine Tabelle, die Spalten für ?obs, ?class und ?label hat. Tabelle 1.3 zeigt das Ergebnis einer derartigen Anfrage gegen das oben aufgeführte Beispieldatenset.

TripleStores implementieren auch die durch OWL oder RDFS zur Verfügung stehenden Inferenzmöglichkeiten und erweitern somit die explizit definierten Daten. Als Technologie für das Semantische Web erfolgt der Zugriff auf TripleStores für gewöhnlich über eine spezielle Ressource, einem sogenannten SPARQL-Endpoint. Für Benutzer präsentiert sich dieser Endpunkt in Form einer speziellen Webseite, auf der man SPARQL-Anfragen eingeben kann. Andere Agenten können direkt über das HTTP-Protokoll zugreifen.

64 Suche bedeutet hierbei, ob es Statements in dem Datenset gibt, auf die das Anfragemuster passt. Durch *?<Zeichen>* wird eine Variable denotiert.

Tabelle 1.3: Ergebnis einer SPARQL-Anfrage.

| ?OBS | ?CLASS | ?LABEL |
|---|---|---|
| Ergebnisse ohne Inferenz | | |
| urn:id234 | urn:LingObs | “e“ |
| Zusätzliche Ergebnisse mit Inferenz | | |
| urn:id234 | qb:Observation | “e“ |
| urn:id235 | qb:Observation | “25℃“ |

**Box 1.6.1 Erklärung des Zustandekommens des Ergebnis der SPARQL Anfrage**

Ohne Inferenz wird nur eine Zeile als Ergebnis produziert, da nur diese Belegung die Triple produziert, die so explizit in dem Datenset stehen. Mit Inferenz werden noch implizit die Aussagen:

```
urn:id234 a qb:Observation .
urn:id235 a qb:Observation .
```

hinzugefügt, die wiederum zwei neue gültige Variablenbelegungen liefern.

TripleStores speichern die in Unterabschnitt 1.5.1 erwähnten RDF-Datensets ab. Es existieren Indizes für alle *Subjekte*, *Objekte*, *Prädikate* und den *Kontext*. Zusätzlich können noch spezielle Indizes zur Freitextsuche oder für geospatiale Anfragen erstellt werden. Graphdatenbanken sind explizit auf die Verwaltung großer Datenmengen ausgelegt. So ist ein effizientes Verwalten von über einer Milliarde Triples für die meisten Graphdatenbanken kein Problem. Moderne Technologien wie Clusterreplikation oder Map Reduce erhöhen diese Zahlen noch weiter.

Für die Analyse der in dieser Arbeit verwendeten Daten ist der TripleStore *GraphDB* der Firma Ontotex[65] ein zentrales Werkzeug. In diesem TripleStore werden die Daten und die verwendeten Ontologien, insbesondere die *phonOntology*, verwaltet. Die Daten werden durch die in der Ontologie definierten Regeln mittels OWL2 RL Inferenz erweitert, so dass ein Laut in seine Lauteigenschaften zerlegt wird.

## 1.7 LINGUISTISCHE ONTOLOGIEN

Als Teil des Semantischen Webs kommen Ontologien bei der automatischen Textanalyse und Klassifizierung zum Einsatz und sind damit auch ein Teilgebiet in der linguistischen Informatik. Sie bieten das Vokabular, um Texte zu analysieren, oder fungieren als lexikalische Datenbank, um Worte zu klassi-

65 <https://ontotext.com>, abgerufen 15.01.2018.

fizieren. Das von der Universität Princeton verwaltete Wordnet[66] (vgl. Fellbaum 2012) ist die bekannteste linguistische Ontologie und diese ist häufig ein Grundbaustein bei dem Verarbeiten natürlichsprachlicher Texte.

Die *Gold-Ontologie* (GOLD)[67] (vgl. Farrar und Langendoen 2003) ist eine umfassende, annotierte und mit Literaturverweisen versehene Ontologie für die Linguistik. Sie versucht die generellen Konzepte der Linguistik zu erfassen sowie abzubilden und kann als eine *Upper-Domain*-Ontologie für die Linguistik betrachtet werden. GOLD versucht auf Basis von Expertenwissen und „best practice"-Erfahrung, eine Wissensbasis für vernetzte linguistische Forschung zu sein. GOLD baut auf der Top-Level-Ontologie SUMO auf und benutzt das Glossar des SIL (Summer Institute of Linguistics)[68] für viele verwendete Termbeschreibungen. Die meisten Konzepte sind annotiert und mit Literaturverweisen versehen. GOLD verwendet den Namensraum *http://purl.org/linguistics/gold/* und wird für gewöhnlich mit dem Präfix *gold:* abgekürzt. Die Hauptunterteilung basiert auf den drei Kernkonzepten von SUMO:

- *gold:Abstract:* Abstrakte Eigenschaften oder Entitäten, die keine eigenständige Ausdehnung in Raum und Zeit haben. Das klassische Beispiel sind Zahlen und mathematische Objekte wie Mengen. Die meisten Konzepte in GOLD sind Unterklassen dieser Klasse. So wird zum Beispiel *gold:Dialect* als Unterklasse *gold:Taxon* aufgefasst mit der Definition:

  > A regional, temporal or social variety of a language, differing in pronunciation, grammar and vocabulary from the standard language,which is in itself a socially favoured dialect. (Hartmann 1973)

- *gold:Object*: „Normale" Dinge in Raum und Zeit. Dies umfasst geographische Regionen, aber auch Menschen oder Gegenstände wie Bücher oder die Orte, an denen ein *gold:Process* stattfindet. In GOLD werden drei Arten von *gold:Object* unterschieden:
  - *gold:SignedLinguisticExpression* zur Repräsentation von Gebärdensprache.
  - *gold:WrittenLinguisticExpression* zur Repräsentation von Schrift.
  - *gold:SpokenLinguisticExpression* zur Repräsentation von gesprochenen Worten[69].

- *gold:Process*: Prozesse sind Entitäten, die temporale Teile haben und nicht vollständig zu einem Zeitpunkt präsent sein können. Beispiele dafür sind: Ein Fußballspiel, Schreiben, Zellteilung oder auch der Vorgang des Sprechens selbst.

Die GOLD-Ontologie beinhaltet insgesamt 503 Klassen und 2918 Axiome, die häufigsten Arten von Axiomen sind dabei Unterklassendefinitionen. Ein

66 <https://wordnet.princeton.edu>, abgerufen 15.01.2018.

67 <http://www.linguistics-ontology.org>, abgerufen 15.01.2018.

68 <https://www.sil.org>, abgerufen 15.01.2018.

69 Hier ist das physikalische „Ding" gemeint, welches durch den Prozess des Sprechens erzeugt wird.

Ausschnitt der Ontologie ist in Abbildung 1.5 zu sehen. Er zeigt einen Teil der Klassen sowie deren Unterklassen und bietet eine grobe Übersicht über die Teilgebiete der Ontologie.


Abbildung 1.5: Ausschnitt aus der GOLD-Ontologie.

GOLD selbst liefert eine Basis für die Eigenschaften der Lauterzeugung. Die Laute selbst werden in GOLD nicht modelliert. Die Eigenschaften, die Unterklassen des *PhoneticProperty*-Astes der Ontologie sind, unterteilen sich nach dem Ort und der Art der Lauterzeugung. Dabei folgen sie den gängigen wissenschaftlichen Annahmen (vgl. Ball u. a. 2008; Crystal 1997). Der Aufbau dieser Taxonomie basiert weitestgehend auf der von Ladefoged (1997). Eine Auflistung der ontologischen Klassen inklusive der dabei verwendeten Definition zu den phonetischen Eigenschaften findet sich in Abschnitt A.1 auf Seite 160. Die Definitionen der Klassen wurden direkt aus der OWL-Datei[70] übernommen. GOLD liegt in einer OWL-Version[71] vor (vgl. Farrar

70 Als Webseitenversion unter: <http://linguistics-ontology.org/gold/2010/PhoneticProperty>, abgerufen 15.01.2018.

71 <http://linguistics-ontology.org/gold-2010.owl>, abgerufen 07.02.2018.

und Langendoen 2009). Die GOLD-Ontologie ist eine Beschreibung linguistischer Eigenschaften und Konzepte und dient damit der Strukturierung sowie als kontrolliertes, interpretierbares Vokabular. Sie ist eine *Upper-Domain*-Ontologie, die für die Entwicklung weiterer Ontologien oder den Informationsaustausch zwischen Agenten dienen kann. Ein direkter Anwendungsbezug ist in GOLD nicht gegeben.

## 1.8 EINE ONTOLOGIE FÜR DIE PHONETIK

Ziel der Ontologie für die Phonetik (*phonOntology*) ist die Möglichkeit aus einer IPA-Notation (Internationales Phonetische Alphabet) auf phonetische Eigenschaften zu schließen und vice versa. Als Anwendungsontologie dient sie dazu, Transkriptionen für eine computergestützte Analyse anwendbar zu machen oder eine Verbindung zwischen leicht formalisierten Sprachdaten und IPA, als normierter Träger phonetischer Informationen, herzustellen. Des Weiteren kann sie als Mediator zwischen verschiedenen Sprachdatenquellen dienen, indem sie Konvertierungsregeln für die unterschiedlichen Ausgangsnotationen liefert. Ein weiterer Aspekt ist die Gewinnung neuer Erkenntnisse über eine Sprachregion durch die Zerlegung der IPA-Phone in ihre Bestandteile. Unter diesem Aspekt wird sie auch in dieser Arbeit eingesetzt (siehe Abschnitt 2.4). Die Ontologie ist in OWL2 implementiert und verwendet die URI <http://issg.de/ontologies/phonetic#> als Namensraum. Als Präfix für diesen Namensraum wird in dieser Arbeit *phon:* verwendet. Das Internationale Phonetische Alphabet[72] gilt als der Ausgangspunkt bei der Entwicklung der *phonOntology* und bedient sich dabei Überlegungen zur hierarchischen Strukturierung phonetischer Eigenschaften, wie sie zum Beispiel von Ladefoged (1988) oder Clements (1985) beschrieben werden. In beiden Publikationen, die wiederum sehr von Chomsky und Halle (1968) inspiriert wurden, wird über eine Strukturierung phonetischer Eigenschaften und der Substitution von Lauten durch eine Matrix von Eigenschaften, die entweder vorhanden (+) oder nicht vorhanden (-) sind, diskutiert. So kann zum Beispiel [m] durch (+voiced, +bilabial, +nasal) ausgedrückt werden. Die Eigenschaften sind wiederum in einer Baumstruktur angeordnet und haben Überklassen, die sie genauer klassifizieren. So kann *Bilabial* als eine Unterklasse von *Place* und *Physiological* beschrieben werden und folgt damit ontologischen Designprinzipien.

Die *phonOntology* implementiert die in IPA definierten Konsonanten und Vokale als Klassen und beschreibt diese Klassen zusätzlich über ihre phonetischen Eigenschaften. Die phonetischen Eigenschaften basieren für die Vokale auf dem Vokaltrapez und bei den Konsonanten auf der Tabelle, die in Abbildung 1.6 dargestellt sind. Diese Eigenschaften decken sich in vielen Fällen mit den Eigenschaften der GOLD-Ontologie, allerdings gibt es auch ein paar Unterschiede, die auf die verschiedenen Zielsetzungen zurückzuführen sind. Während GOLD einen Fokus auf die Konfiguration des Mundraumes setzt, liegt der Fokus der *phonOntology* auf der Konfiguration der Lauterzeugung.

72 <https://www.internationalphoneticassociation.org>, abgerufen 07.02.2018.

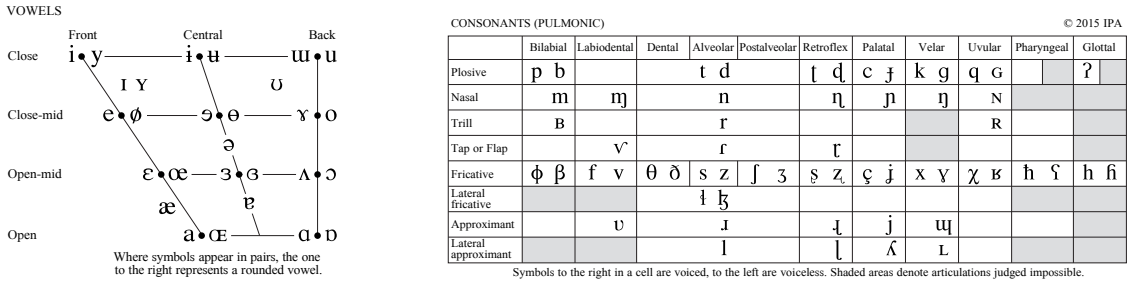



Abbildung 1.6: IPA Chart, <http://www.internationalphoneticassociation.org/content/ipa-chart>, verfügbar gemacht unter der Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2015 International Phonetic Association.

In seinen Arbeiten zur Onlineplattform PHOIBLE[73] (vgl. Moran 2012; Moran, McCloy und Wright 2014) stellt Steven Moran ein ähnliches System vor wie die *phonOntology*. Der Fokus dabei liegt auf der Sammlung und Vernetzung der verschiedenen phonologischen Lautinventare weltweit und bietet eine umfassende Datenbank an IPA-Lauten und deren Vorkommen in Sprachen, sowie eine Annotierung dieser Laute durch die bereits erwähnte [+/-]-Eigenschaftsmatrix[74]. Durch die Verwendung von IPA und einer Implementierung in OWL kann die *phonOntology* leicht mit dem PHOIBLE-System verknüpft werden. Dies bietet der *phonOntology* einen guten Einstiegspunkt in die *Linguistic Linked Open Data Cloud*[75] (vgl. Chiarcos, Nordhoff und Hellmann 2012), einem System basierend auf dem Semantischen Web, um linguistische Daten verfügbar und untereinander kompatibel zu machen.

Die benannten Laute der *phonOntology* sind unterteilt in Konsonanten und Vokale. Zusätzlich gibt es noch den „Dummy"-Laut *Gap*, der den Ausfall eines Lautes repräsentiert. Jeder benannte Laut ist zudem über die vordefinierte OWL-Relation *owl:equivalentClass* mit einer Kombination von phonetischen Eigenschaften verbunden. In dieser Hinsicht kann die *phonOntology* als eine Implementierung der oben erwähnten Eigenschaftsstrukturierungen gesehen werden, aber anstelle einer Definition über eine [+/-]-Matrix der Eigenschaftsausprägungen erfolgt die Implementierung als ontologische Axiome in OWL2. So wird zum Beispiel das Phon [ɪː] in der Ontologie durch folgendes Axiom ausgedrückt:

```
Vowel
 and (vowelBackness value NearFront)
 and (vowelAperture value LoweredClose)
 and (vowelLongness value Long)
 and (vowelRoundness value Unround)
```

Umgangssprachlich bedeutet dies, dass ein [ɪː]-Laut ein Vokal ist, der die Vokalöffnungsgradeigenschaft (*Backness*) *NearFront*, die Vokalhöheneigenschaft (*Aperture*) *LoweredClose*, die Vokallängeneigenschaft (*Longness*) *Long*

73 <https://phoible.org>, abgerufen 23.09.2018.
74 Im Kontext der Phonologie siehe: Hayes (2011).
75 Siehe dazu: <http://linguistic-lod.org>, abgerufen 23.09.2018.

und die Vokalrundungseigenschaft (*Roundness*) *Unround* hat. Die Eigenschaften werden durch URIs repräsentiert und sind die Stellvertreter für die in IPA verwendeten Konzepte der Lautbildung. Somit repräsentiert die URI *<http://issg.de/ontologies/phonetic#LoweredClose>* die fast geschlossene Mundöffnung in der vokalischen Lauterzeugung. Die phonetischen Eigenschaften sind hierarchisch angeordnet. So teilen sie sich zunächst in konsonantische und vokalische Eigenschaften und die „Dummy"-Eigenschaft *Nil* auf, welches die Lautausfall-Eigenschaft für den *Gap*-Laut darstellt. Die konsonantischen Eigenschaften unterteilen sich weiter in Art, Ort und Stimmhaftigkeit der Artikulation. Die Vokaleigenschaften sind unterteilt in Öffnungsgrad, Zungenposition, Lautlänge und Lippenrundung.

Die Diphthongeigenschaften werden gesondert behandelt. Strukturell sind sie aber eine Unterklasse der vokalischen Lauteigenschaften. Sie werden durch einen Anfangs- und einen Endlaut repräsentiert und nicht durch eine direkte Konfiguration im Vokaltrapez. Die Diphthonglauteigenschaften sind dadurch bereits eine spezifische Konfiguration und die Repräsentation eines Diphthongs erfolgt über zwei dieser Lauteigenschafteskonfigurationen.

Der *PhoneticProperty*-Ast der Ontologie ist in Tabelle 1.4 aufgelistet. Die hervorgehobenen Eigenschaften haben noch Unterklassen, die einfachen Eigenschaften sind Blätter im Baum[76]. Damit ist die Eigenschaft *Nil* ein Blatt der Hauptkategorie *PhoneticProperty*. Die beiden Tonakzente (TA1, TA2) (siehe Seite 43) sind Blätter der *Intonations*-Oberklasse, die wiederum eine Unterklasse der *PhoneticProperty* ist. Zu diesen Klassen gibt es noch passende Relationen, die alle eine Unterklasse von *phoneticProperty*[77] sind. Die Relationen sind:

- *phoneticProperty*: Die Hauptrelation für alle genaueren phonetischen Eigenschaftsrelationen; hat als *rdfs:range* eine *PhoneticProperty*
  - *consonantProperty*: Die übergeordnete Relation für konsonantische Eigenschaften.
    - *articulationManner*: Ordnet einem konsonantischen Laut eine Artikulationsart zu.
    - *articulationPhonation*: Ordnet einem konsonantischen Laut die Art der Stimmhaftigkeit zu.
    - *articulationPlace*: Ordnet einem konsonantischen Laut den Ort der Lauterzeugung zu.
  - *prosodicProperty*: Die Relation, um Lauten prosodische Eigenschaften zuzuordnen. Dies ist eine Erweiterung zu IPA, die notwendig ist, um die Tonakzente in die Analyse in Kapitel 4 miteinzubeziehen.
  - *vowelProperty*: Die übergeordnete Relation für vokalische Eigenschaften.
    - *diphthongProperty*: Die übergeordnete Relation für die Diphthongeigenschaften. Diese Relation hat hauptsächlich struk-

76 In der Graphentheorie werden die Endknoten in einem Baum Blätter genannt.

77 Eine verbreitete Konvention bei der Ontologieerstellung ist, dass Relationen mit Kleinbuchstaben und Klassen mit Großbuchstaben anfangen.

turierenden Charakter und sollte nicht direkt verwendet werden.

  - *diphthongStart*: Die Relation, die einen Diphthong mit der Anfangskonfiguration seiner Artikulation verbindet.
  - *diphthongEnd*: Die Relation, die einen Diphthong mit der Endkonfiguration seiner Artikulation verbindet.

* *monophthongProperty*: Die übergeordnete Relation für die Monophthongeigenschaften.
  - *vowelAperture*: Die Relation, die einen vokalischen Laut mit einer Öffnungsgradeigenschaft verbindet.
  - *vowelBackness*: Die Relation, die einen vokalischen Laut mit einer Zungenpositioneigenschaft verbindet.
  - *vowelLongness*: Die Relation, die einen vokalischen Laut mit einer Längeneigenschaft verbindet.
  - *vowelRoundness*: Die Relation, die einen vokalischen Laut mit einer Lippenrundungseigenschaft verbindet.

Ein in IPA benannter Laut kann nun über die *owl:equivalentClass* Relation mit den entsprechenden Eigenschaften verbunden werden. Die Kombination der Lauteigenschaften bildet eine anonyme Klasse. Da nur wenig explizite Verbote implementiert sind, ist es theoretisch möglich, Laute, die entweder physikalisch nicht umsetzbar oder nicht explizit definiert sind, über diese anonymen Klasse zu definieren. Auch ist es möglich, anonyme Lautklassen unter bestimmten Gesichtspunkten zu definieren. So kann zum Beispiel die Lautklasse aller Vorderzungenvokale erzeugt werden als :

```
Vowel and (vowelBackness value Front)
```

Dies geschieht gewöhnlich implizit durch die Inferenz, die durch das OWL2 RL Profil definiert ist. Die Hauptstärke der Inferenz basiert allerdings auf dem Zerlegen der IPA-Laute in Lauteigenschaften. Durch die in OWL2 definierten *owl:equivalentClass*-Relation lässt sich eine Verbindung zwischen einem IPA-Laut und der dazugehörigen anonymen Eigenschaftsklasse herstellen. So lässt sich der IPA-Laut als Repräsentant dieser Eigenschaften auffassen. In einer Anwendung bedeutet das, dass Instanzen eines IPA-Lautes automatisch die entsprechenden Lauteigenschaften zugeordnet bekommen. Da IPA ein weitestgehend normiertes Format ist, lassen sich so semiautomatisch[78] phonetische Observationen mittels der *phonOntology* und OWL2 RL Inferenz in ihre Lauteigenschaften zerlegen, die als Basis für weiterführende Auswertungen dienen können.

In der *phonOntology* werden 65 Vokale und 64 Konsonanten explizit aufgeführt. Diese Laute werden durch URIs aus dem *phon:* Namensraum repräsentiert (vgl. Seite 39). Der Identifikator eines Lautes folgt den offiziellen

78 Semiautomatisch, da die Zuordnung der Observationen zu entsprechenden IPA-Lauten noch manuellen Input benötigt. Mehr dazu in Abschnitt 2.4.

IPA Bezeichnungen[79]. Bei den Vokalen wird noch eine explizite Unterscheidung zwischen kurz und lang vorgenommen. Das Phon [ɪː] wird zum Beispiel durch die Ressource *phon:LongLoweredCloseNearFrontUnroundedVowel* repräsentiert.

Insgesamt besteht die *phonOntology* aus 5665 Axiomen, die sich aufteilen in 200 Klassendefinitionen, 15 Relationendefinitionen und 48 Individuendefinitionen. Alle übrigen Axiome beschreiben die Beziehungen zwischen diesen.

*Abweichungen gegenüber IPA*

Auch wenn es Ziel der *phonOntology* ist, die Lautdefinitionen aus IPA so gut wie möglich umzusetzen, gibt es doch ein paar Abweichungen. Eine Abweichung ist der deutsche *a*-Laut. Zusätzlich zu den *a*-Varianten auf der Öffnungsgrad-Achse des Vokaltrapezes wird noch ein zentraler *a*-Laut hinzugefügt (vgl. International Phonetic Association 1999, S. 87; Hall 2011, S. 34; Pompino-Marschall 2009). Die Bezeichnung in der Ontologie ist *OpenCentralUnroundedVowel* beziehungsweise *LongOpenCentralUnroundedVowel* für die lange Ausprägung. Die Denotationen in IPA für diesen Laut sind [a̟] und [a̟ː] für den entsprechenden Langvokal. Eine weitere Abweichung ist das Hinzufügen von zwei Intonationsmerkmalen, den sogenannten Tonakzent 1 (TA1) („Schärfung“) und Tonakzent 2 (TA2) („Trägheitsakzent“). Die Tonakzente sind ein wichtiger Bestandteil des Untersuchungsgegenstands dieser Arbeit und basieren auf der Arbeit von Schmidt (1986). Hier werden sie als zusätzliche phonetische Eigenschaften angesehen und sind den anderen Eigenschaften gegenüber gleichgestellt. Außerdem wird ein Leer-Laut hinzugefügt, einmal als definierter Laut an sich (*Gap*), aber auch als entsprechende Eigenschaft, die mit *Nil* denotiert wird. Damit wird der Wegfall eines Lautes markiert. Dies ist nötig, da der Wegfall eines Lautes eine wichtige Observation ist (vgl. Tabelle 2.3), Ontologien aber auf der Offene-Welt-Annahme und monotoner Logik basieren und man deshalb einen Wegfall explizit modellieren muss. Diese zusätzlichen Merkmale wurden hinzugefügt, um eine bessere Transformation der in Kapitel 2 vorgestellten Daten zu ermöglichen.

*Probleme und Grenzen*

Eine Ontologie kann immer als „work in progress“ angesehen werden, da sich mit neuen Erkenntnissen über eine Domäne auch deren ontologischen Annahmen ändern können. Ebenso sind Anwendungsontologien wie die *phonOntology* fokussiert auf eine Anwendung und müssen nicht notwendigerweise eine Domäne komplett abdecken. Diese Herausforderung ist den Domänenontologien wie GOLD vorbehalten, meistens auf Kosten der Anwendbarkeit. Der Fokus der *phonOntology* liegt auf der Transformation der IPA-Laute in ihre Lauteigenschaften, um so ein statistisches Modell über diese Eigenschaften für eine Region aufstellen zu können. Um diese Aufgabe anwendbar und überschaubar zu halten, gibt es ein paar Limitationen.

79 Zu finden unter <https://www.internationalphoneticassociation.org/content/ipa-chart>, abgerufen 16.01.2018.

So werden bis auf die Länge bei Vokalen keine Diakritika aus IPA berücksichtigt. An expliziten Lauten sind nur die Basislaute aufgeführt. Diakritika erweitern die Menge der Laute beträchtlich, machen es aber auch unpraktikabel, alle möglichen Laut-Diakritika-Kombinationen explizit zu definieren. So verfügen wir selbst nur über ein überschaubares Basisvokabular an IPA-Lauten und erzeugen komplexere Laute durch die Kombination mit Diakritika. So steht [ë:] für ein langes, zentralisiertes [e]. Das [e] selbst ist wiederum ein Surrogat für die in IPA definierten Eigenschaften *OpenMid, Front, Unround*[80]. In der Ontologie lassen sich Eigenschaftskombinationen, die nicht einem expliziten Laut zugeordnet sind, als anonyme Klasse ausdrücken, die durch eine Blank Node repräsentiert werden. Für einen Menschen trägt eine solche Blank Node keine interpretierbare Information mehr, für den Computer stellt sie aber kein Problem dar. Diakritika können damit als zusätzliche Eigenschaften zu einem Laut angesehen werden, und die Ontologie kann leicht um diese zusätzlich benötigten Eigenschaften erweitert werden. Dies wirft die Frage auf, wie diese Eigenschaften einzuordnen sind und in welchem Verhältnis diese Eigenschaften zu den Haupteigenschaften stehen. Bei manchen Eigenschaften ist der Fall einfach. So kann eine Aspiration als zusätzliche phonetische Eigenschaft hinzugefügt werden und die ursprüngliche Eigenschaftsmenge des aspirierten Lautes wird um dieses neue Merkmal erweitert. Bei anderen Diakritika ist dies nicht so einfach möglich. Ist zum Beispiel die Fortisierung des [d] in IPA denotiert durch [d̥] gleichzusetzen mit einem [t] oder einem [t̬]-Laut? In IPA markieren die Symbole [◌̥] und [◌̬] die Eigenschaften *voiceless* und *voiced*, in der Praxis werden sie aber zur Beschreibung von Fortis und Lenis eingesetzt, die Druckeigenschaften bei der Lauterzeugung markieren und mit stimmlos oder stimmhaft korrelieren (vgl Kohler 1984). Ontologien sind kategorienbasiert, was eine gewisse Distinktivität impliziert. Lauteerzeugung ist aber ein Prozess, das heißt, ein Laut kann nie zu einem definierten Zeitpunkt existieren, sondern benötigt immer eine zeitliche Ausdehnung. Auch ist Lauterzeugung nicht deterministisch, sondern bildet ein Spektrum an Lauten. Dies ermöglicht nicht immer eine saubere Trennung in zwei disjunkte Klassen. Ein IPA-Symbol für einen Laut repräsentiert damit nicht genau eine Lautkonfiguration, sondern ein Spektrum an Konfigurationen, wobei die Grenzbereiche nicht immer eindeutig oder einfach zu bestimmen sind. Eine Notation in IPA ist eine Interpretation oder Klassifikation eines Lauterzeugungsprozesses und ist damit bereits einer Glättung unterworfen, die zum Beispiel durch das zur Verfügung stehende Glypheninventar oder durch die Kombinationsmöglichkeiten von Basis- und Diakritikaglyphen beschränkt ist (vgl. Moran und Cysouw 2018).

Moderne Ansätze zur automatischen Lautklassifikation bieten oft Möglichkeiten für eine gewichtete oder wahrscheinlichkeitsbasierte Klassifikation anstelle einer einfachen Lautzuordnung (vgl. Graves und Jaitly 2014; Keil 2017; Pellegrini und Mouysset 2016). Diese Methoden stehen allerdings für diese Arbeit nicht zur Verfügung und bieten auch keine vollständige Lösung für dieses Problem. Je nach Genauigkeit (Granularität) der Ontologie müssen Entscheidungen getroffen werden, zu welcher Klasse ein Laut zugeordnet wird. Dieses Problem tritt bereits in der Notation der Ausgangsdaten in IPA

80 Ohne explizite zusätzliche Kennung wird ein Laut als *Short* angenommen.

auf, wird aber häufig durch Zusatzbemerkungen oder zusätzliche Diakritika abgeschwächt.

Die *phonOntology* versucht so weit möglich, mit den Haupteigenschaften und einigen auf das Datenset des *Mittelrheinischen Sprachatlas* zugeschnittenen Erweiterungen auszukommen. So können die Tonakzente als zusätzliche Eigenschaft erfasst werden und einem Laut können mehrere Artikulationsorte zugewiesen werden, um Unschärfe zwischen zwei Merkmalen abzubilden.

Auch bei einem Mapping zwischen der GOLD-Ontologie und der *phonOntology* müssen einige Eigenheiten beachtet werden. So werden die Lauteigenschaften in der GOLD-Ontologie unter einem anatomischen Gesichtspunkt beschrieben, wohingegen die *phonOntology* ergebnisorientiert ist. Damit ist gemeint, dass die phonetischen Eigenschaften so gewählt sind, dass damit IPA-Laute erzeugt werden können. So unterscheidet die *phonOntology* sieben Öffnungsgrade bei der Vokalerzeugung, die GOLD-Ontologie aber nur drei. Genauere Unterteilungen erfolgen über andere Eigenschaften. Das bedeutet, dass keine einfache 1:1-Translation zwischen diesen Ontologien möglich ist und in vielen Fällen ein Mapping über komplexe Klassen zu erstellen ist.

Tabelle 1.4: Auflistung aller phonetischen Eigenschaften in der *phonOntology*.

| **PhoneticProperty** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ConsonantArticulation** | | | Nil | **Intonation** | | **VowelConfiguration** | | | | | |
| **ArticulationManner** | **ArticulationPhonation** | **ArticulationPlace** | | TA1 | TA2 | **Aperture** | **Backness** | **DiphthongConfiguration** | | **Longness** | **Roundness** |
| Affiricate | Voiced | Alveolar | | | | Close | Back | DiphthongEnd | DiphthongStart | ExtraShort | Round |
| Approximant | Voiceless | Bilabial | | | | CloseMid | Central | DiphLoweredCloseNearBack | DiphOpenCentral | ExtraLong | Unround |
| Flap | | Dental | | | | LoweredClose | Front | DiphLoweredCloseNearFront | DiphOpenMidBack | HalfLong | |
| Fricative | | Glottis | | | | Mid | NearBack | | DiphOpenMidFront | Long | |
| LateralApproximant | | Labiodental | | | | Open | NearFront | | | Short | |
| Nasal | | Palatal | | | | OpenMid | | | | | |
| Plosive | | Pharyngeal | | | | RaisedOpen | | | | | |
| Trill | | Postalveolar | | | | | | | | | |
| | | Retrofelx | | | | | | | | | |
| | | Uvular | | | | | | | | | |
| | | Velar | | | | | | | | | |

# 2 DER MITTELRHEINISCHE SPRACHATLAS

Der *Mittelrheinische Sprachatlas* (MRhSA) (vgl. Bellmann, Herrgen und Schmidt 1994–2002) dient als Anwendungsfall für die *phonOntology*. Die Daten, die für den Atlas erhoben wurden, sind in der Onlineplattform REDE[81] (vgl. Schmidt, Herrgen und Kehrein 2008b) integriert und bilden die Grundlage der Clusteranalyse in Kapitel 4. Dieses Kapitel bietet einen Überblick über die Methodik und die Ergebnisse des MRhSA und die Aufbereitung der Daten mittels eines TripleStores und der Anwendung der *phonOntology* als vorbereitenden Schritt zur Clusteranalyse.

## 2.1 DER MRHSA - EIN ÜBERBLICK

Der MRhSA ist einer der ersten bidimensionalen Sprachatlanten und der erste europäische. Zudem gilt er als ein Ausgangsmodell für nachfolgende bibeziehungsweise pluridimensionale Sprachatlanten oder linguistische Kartierungsprojekte wie dem *L'Atlas Linguistique Diatopique et Diastratique de l'Uruguay* (ADDU) von Thun (2001) oder der *Neuerhebung der modernen Regionalsprachen des Deutschen* von Schmidt, Herrgen und Kehrein (2008a). Neben der sprachlichen Dimension wird im MRhSA auch eine soziologische Dimension berücksichtigt. Es werden also nicht nur die reinen Sprachdaten gesammelt, sondern es wird auch der soziale Hintergrund der Informanten berücksichtigt. Diese Methodik wurde zuerst von Fujiwara (1976) und Kurath (1973) angewendet und bildete die Grundlage für moderne sprachdynamische Analysen (vgl. Schmidt und Herrgen 2011, S. 145). Als soziologische Dimensionen spielen Alter und Ortsfestigkeit eine besondere Rolle. Der Atlas bietet zwei Datensets unter Berücksichtigung dieser Dimensionen. Zum einen die „ältere Generation" (Datenserie 1), die für ortsgebundene Personen um die 75 Jahre steht, und zum anderen die „jüngere Generation" (Datenserie 2), die für mobile[82] Personen um die 35 Jahre zur Zeit der Datenerhebung steht. Für beide Gruppen gilt, dass die Familien seit mindestens zwei Generationen ortsansässig und im handwerklichen oder landwirtschaftlichen Bereich tätig sind. Zusätzlich wird noch der Bildungsgrad als Dimension einbezogen[83].

Das Vorhandensein von zwei Generationen bietet die Möglichkeit einer „apparent time"-Analyse (vgl. Labov 1994), da neben einer generationeninternen Analyse auch der Vergleich zwischen den beiden Generationen zu einem Zeitpunkt möglich ist.

Die Datenerhebung für Datenserie 1 erfolgte an 549 ausgewählten Orten des Untersuchungsgebietes[84] mit insgesamt 1680 Informanten. Für die Erhebung der jüngeren Generation wurden die Orte auf 292 und insgesamt

81 <https://www.regionalsprache.de>, abgerufen 15.02.2018.

82 Mobil bedeutet hier Nahpendler. Insgesamt wird auf eine gewisse Ortsfestigkeit geachtet.

83 Diese Dimension wird in dieser Arbeit aber nicht weiter berücksichtigt.

84 Das Untersuchungsgebiet wird in Abschnitt 2.5 auf Seite 58 genauer vorgestellt.

830 Informanten reduziert, wobei auf eine ausgeglichene Verteilung in der Fläche des Untersuchungsgebietes geachtet wurde. Aus der Datenerhebung gingen 1073 Karten hervor, die sich auf fünf Bände aufteilen. Dabei wurde zu jedem untersuchten Phänomen entweder eine (nur Datenserie 1) oder drei Karte(n) (Datenserie 1, Datenserie 1 eingeschränkt auf die Orte zur Datenserie 2, Differenzkarte aus Datenserie 1 und 2, in der Unterschiede in der Serie 2 rot hervorgehoben sind) erstellt. Die Bände selbst sind unterteilt nach[85]:

Bd. 1: Vorkarten; Diphthonge des Bezugssystems mit 13 Vorkarten und 115 Sprachkartenblätter zu den Diphthongen des Mittelhochdeutschen. Publiziert 1994.

Bd. 2: Langvokale des Bezugssystems mit 135 Sprachkarten zum mittelhochdeutschen Langvokalismus. Publiziert 1995.

Bd. 3: Kurzvokale des Bezugssystems mit 219 Karten zum mittelhochdeutschen Kurzvokalismus. Publiziert 1997.

Bd. 4: Vokale in Nebensilben; Konsonanten des Bezugssystems mit 272 Sprachkarten zum westgermanischen Konsonantismus. Publiziert 1999.

Bd. 5: Morphologie mit zwei Vorkarten und 309 Karten zur Morphologie. Publiziert 2002.

Die Karten sind hauptsächlich Punkt-Symbol-Karten im Maßstab 1:60000 für die ältere Generation und 1:1000000 für die vergleichenden Kontrastkarten. Die Datenerhebung erfolgte von 1978 bis 1988 und der Bearbeitungszeitraum des MRhSA war von 1983 bis 2001 (vgl. Girnth 2015, S. 30).

Die Datenerhebung erfasst phonologische und morphologische Phänomene. In dieser Arbeit werden aber nur erstere berücksichtigt. Die Phänomene wurden direkt vor Ort während der Erhebung notiert, dadurch konnte bei Unsicherheit (siehe Abschnitt 2.2) direkt nachgefragt werden. Als Notationssytem für die Erhebung dient ein ausgewähltes Vokabular aus IPA. Dies erlaubt zum einen eine Vergleichbarkeit mit anderen Atlanten oder Datenerhebungen, die auf IPA basieren und gewährleistet zum anderen eine Normalisierung während der Erhebung (vgl. Bellmann 1994, S. 88 ff.). Die Wahl von IPA als Notationssytem ist einer der Gründe, warum der MRhSA für eine ontologiebasierte Clusteranalyse geeignet ist. Zusätzlich zur direkten Notation wurden alle Interviews auf Tonband aufgezeichnet. Dies erlaubt eine nachträgliche Überprüfung und falls notwendig Anpassung beziehungsweise Angleichung der Notationen. Angleichungen sind zum Beispiel notwendig, wenn Exploratoren individuelle Variation für die Notation eines Phänomens aufweisen. Um diesen Umstand von vornherein zu reduzieren, wurden bei der Erhebung zwei speziell ausgebildete Hauptexploratoren eingesetzt, die außerdem alle erhobenen Daten überprüften.

85 <https://www.igl.uni-mainz.de/forschung/dialektforschung/mittelrheinischer-sprachatlas.html>, abgerufen 16.01.2018.

Die erhobenen Daten sind nach historischen Lautbezugssystemen klassifiziert. Dies ermöglicht das Vergleichen von räumlich verteilten linguistischen Phänomenen auf Basis einer gemeinsamen, historischen Grundlage.

*Die historischen Bezugssysteme des MRhSA*

Die historischen Bezugssysteme basieren auf der Annahme, dass Laute einen gemeinsamen historischen Bezugspunkt haben und so zu einer (Laut-) Klasse zusammengefasst werden können. Dies kann im ontologischen Sinne als Klassen-Instanz-Beziehung gesehen werden, bei dem die Worte die Instanzen zu den historischen Lautklassen bilden[86].

Für den Vokalismus wurde das HISTORISCHE MITTELHOCHDEUTSCH und für den Konsonantismus das HISTORISCHE WESTGERMANISCHE als Bezugssystem gewählt. Das HISTORISCHE MITTELHOCHDEUTSCH basiert auf den Arbeiten von Karl Lachmann (1793–1851) und galt bei der Ausarbeitung des MRhSA als die Bezugsnorm, die der Dialekt der Region im 10. bis 13. Jahrhundert durchlaufen hat (vgl. Wiesinger 1983, S. 1045).

Neuere Forschungen, auch basierend auf den Ergebnissen des MRhSA, ziehen die Anwendung des HISTORISCHEN MITTELHOCHDEUTSCH als Bezugssystem für den Vokalismus für die gesamte Region in Zweifel und schlagen stattdessen für das MOSELFRÄNKISCHE (siehe Abbildung 2.3) ein angepasstes Lautsystem (ALTWESTDEUTSCH) für den Vokalismus vor (vgl. Schmidt 2015), das sich direkt aus einem regional angepassten Westgermanisch ableitet.

Das HISTORISCHE WESTGERMANISCHE als Bezugssystem für den Konsonantismus gilt als deutlich sicherere Annahme, da sich das Untersuchungsgebiet durch die Hauptgrenze der zweiten Lautverschiebung einteilen lässt. Diese Grenze, die vornehmlich durch die *dat/das*- und *Dorp/Dorf*-Isoglosse bestimmt wird, ist eine der bestimmenden sprachlichen Unterscheidungsmerkmale in dieser Region.

Da die Korrektheit dieser Bezugssysteme sich nicht immer vollständig belegen lässt, werden die Systeme eher als Ordnungs- oder Markierungssysteme gesehen, in denen eine ähnliche Lauterwartung an die Instanzen einer Wortklasse gesetzt wird. In diesem Sinne werden auch die Bezugssysteme bei der Clusteranalyse in Kapitel 4 betrachtet. Tatsächlich ist die Analyse der Varianz innerhalb einer Klasse ein zentraler Punkt der Clusteranalyse. Eine vollständige historische Korrektheit ist dabei nebensächlich. Die beiden Bezugssysteme wurden direkt aus den Klassifikationen entnommen, die den Karten zugeordnet sind. Dies bedeutet, dass, mit Ausnahme der Hauptanalyse, die alle Laute unabhängig des historischen Bezugssystems mit einbezieht, die ausgewiesenen Cluster in der Clusteranalyse abhängig von dem gewählten Bezugssystem sind.

Der *Sprachatlas des Deutschen Reichs*[87] (vgl. Wenker 1877; Wenker und Wrede 1888–1923) untersucht Wörter, die als Instanzen oder Referenzwör-

86 In gewissem Sinne wird eine zusätzliche Metaebene eingeführt. Das „was“ zum Beispiel realisiert als [vas̯], welches von einer bestimmten Person ausgesprochen wird, kann als Instanz der Klasse *was* gesehen werden. Dieses *was* wiederum kann als eine Instanz der Bezugsklasse *wg. t* gesehen werden. In der Ontologieentwicklung nennt man diese Methodik „punning“ (vgl. Grau u. a. 2008).

87 Die bereits erwähnte Wenkererhebung. Oft als WA (Wenkeratlas) abgekürzt.

ter der historischen Lautklassen fungieren. So dient die Wenkererhebung als Ausgangs- oder Referenzpunkt für viele Sprachatlanten. Diese Wörter sind Teil der sogenannten „Wenkersätze", einer Sammlung an speziell konstruierten Sätzen, mit der ein breites Spektrum dialektaler Phänomene erfasst wurde. Eine Auflistung der im MRhSA verwendeten historischen Bezugsklassen mit zugehörigen standardsprachlichen Referenzwörtern findet sich in Abschnitt A.3. Diese Kompatibilität mit der Wenkererhebung ermöglicht ein leichtes Einbinden des MRhSA in eine „real-time"-Analyse (vgl. Labov 1994). Dadurch, dass dieselben Phänomene annotiert wurden, lässt sich ein direkter Vergleich zwischen distinkten Zeitabschnitten vornehmen. Die Möglichkeit innerhalb des Atlas bereits zwei Generationen zu kontrastieren und die Kompatibilität mit der Wenkererhebung machen den MRhSA zu einem sprachdynamischen Atlas (vgl. Schmidt und Herrgen 2011, S. 145 f.).

Da der MRhSA mit dem verwendeten Bezugssystem eine Klassifikation der Laute vornimmt, die Daten ein hohes Maß an Normierung in Form des IPA-Vokabulars aufweisen und die Daten bereits digitalisiert und in eine Datenbank überführt wurden, bietet sich dieser Atlas als geeigneter Datensatz an, um durch die *phonOntology* annotiert zu werden.

## 2.2 FEHLER UND FEHLERQUELLEN

Da diese Arbeit auf sekundären Daten aufbaut, ist es wichtig, mögliche Fehler zu berücksichtigen. Mögliche Fehlerquellen sind die Atlanten selbst und diese werden bereits in dem Einführungsband zum MRhSA (siehe Bellmann 1994, S. 123 ff.) behandelt. Diese Fehlerquellen sind unterteilt in:

PRODUKTIONSFEHLER beschreiben Fehler oder Anomalien bei der Lautproduktion. Diese Fehlerquelle liegt damit beim Informanten selbst. Da dieser Fehler durch den individuellen Sprechapperat des Informanten zustande kommt, lässt er sich nur schwer vermeiden. Mehr Informanten für einen Ort und eine hohe Varianz an Wortkombinationen helfen diesen Fehler zu reduzieren.

VERSTEHENSFEHLER sind Fehler, die durch Missinterpretation der Fragen zustande kommen. Diese Art von Fehler lassen sich durch Nachfragen oder genauere Erklärung der Frage reduzieren.

HYPEDIALEKTALISMEN beschreiben dialektale Übertreibungen oder Korrekturen, die durch Unsicherheit oder Stress zustande kommen können. Ähnlich wie die PRODUKTIONSFEHLER helfen mehr Informanten oder Lauterzeugung in unterschiedlichen Kontexten bei der Minimierung dieses Fehlers.

NOTIERUNGSFEHLER sind Fehler des Explorators während der Datenerfassung. Da die Datenerfassung in den meisten Fällen direkt während des Interviews erfolgte, kann es natürlich zu solchen „Flüchtigkeitsfehlern" kommen. Zusätzliche Audioaufnahmen und eine doppelte Überprüfung durch die Hauptexploratoren helfen diesen Fehler zu minimieren.

KARTIERUNGSFEHLER sind Translationsfehler beim Überführen der Daten auf die Karte. Das Übertragen der Daten auf Karten erfolgte manuell, was bedeutet, dass auch dort Flüchtigkeitsfehler entstehen können. Doppelte Überprüfung hilft auch hier wieder bei der Minimierung.

Bevor die Daten für die Clusteranalyse bereit stehen, durchlaufen sie noch drei weitere Schritte, in denen wiederum Fehler auftreten können. So können bei der Integration der Daten in die Datenbank des REDE-Projekts Übertragungsfehler[88] vorkommen, genauso können wiederum Übertragungsfehler bei der Transformation des relationalen Datensatzes in REDE in das RDF-Datenset und der anschließenden Integration in die Graphdatenbank vorkommen. Eine letzte Fehlerquelle ist das Mapping zur *phonOntology*, dies geschieht semiautomatisch mithilfe einer Mappingtabelle und wird genauer in Abschnitt 2.4 behandelt. Die aufgelisteten Fehlerquellen wurden bereits während der Erhebung kontrolliert und auch die Integration in das REDE-Projekt erfolgte kontrolliert. Zudem sind diese Daten bereits seit Jahren in aktiver Verwendung und damit wiederholter Überprüfung unterzogen, so dass mögliche Fehler allenfalls in den letzten beiden Schritten zu vermuten sind. Da die Transformation in ein RDF-Datenset automatisch erfolgt, ist zu erwarten, dass diese Transformationsfehler systematisch sind. Stichproben haben keine unerwarteten Differenzen zwischen den Daten in REDE und den transformierten Daten des RDF-Datensets ergeben. Auch wenn nicht davon auszugehen ist, dass ein grober systematischer Fehler während des ETL-Prozesses von den Karten bis hin zum RDF-Datenset aufgetreten ist, so ist doch wichtig anzumerken, dass das RDF-Datenset nur eine Approximation des Sprachraumes ist. An mindestens zwei Stellen wurde bereits eine Datenglättung[89] vorgenommen. Zum einen bildet die *phonOntology* nicht das vollständige Spektrum der menschlichen Lauterzeugung ab, sondern beschränkt sich (mit einigen Ausnahmen) auf die in IPA definierten Hauptlaute. Zum anderen ist das Mapping zwischen den Lautnotationen des MRhSA und der *phonOntology* nicht eineindeutig. Da eine Clusteranalyse per Definition einen Fokus auf Makrostrukturen setzt und bereits in der Vorbereitung des Datensets gewisse Glättungsprozesse (siehe Abschnitt 3.2) stattfinden und quantitative Merkmale im Fokus stehen, sollten sich die Auswirkungen dieser Fehler in Grenzen halten[90].

88 Mit Übertragungsfehler sind nicht Fehler im Sinne einer fehlerhaften Datenübertragung bei einer Netzwerkkommunikation gemeint, sondern Fehler, die während des ETL-Prozesses (**E**xtract, **T**ransform, **L**oad) auftreten.

89 Datenglättung beschreibt den Vorgang einer Vereinfachung der Daten, indem zum Beispiel seltene Sonderfälle (Ausreißer) ignoriert werden oder komplexe Datenstrukturen in einfachere überführt werden. Dabei geht immer etwas Information verloren. Dieser Fehler ist aber für gewöhnlich Teil einer Datentransformation und lässt sich kaum vermeiden. Ziel sollte es natürlich sein, diesen Fehler in einem angemessenen Maße zu minimieren.

90 Während der iterativen Entwicklung der Framework zur Clusteranalyse haben sich keine nennenswerten Unterschiede in den Ergebnissen, basierend auf leicht angepasste Datensets, ergeben.

## 2.3 DER MRHSA IN REDE

Der *Mittelrheinische Sprachatlas* wurde im Zuge des REDE-Projekts[91] digitalisiert[92]. Dazu wurden die Daten in den Karten in eine relationale Datenbank überführt. Zudem wurden ausgewählte Karten georeferenziert und als Rasterkarten verfügbar gemacht. Diese Konvertierung der Karten erfolgte semiautomatisch aus den HPGL-Dateien[93], die den Karten zugrunde liegen. Dabei wurde versucht, so viele Informationen wie möglich aus den zugrundeliegenden Datensätzen zu extrahieren und die fehlenden Informationen nachträglich manuell einzugeben. Auch wurde darauf geachtet, dass die Karteninformationen in Form von Symbolzuordnungen an Orten mit entsprechender Erklärung dieser Symbole in einer Legende auch in der Onlineanwendung erhalten bleibt. Diese Daten können dann in Anwendungen wie dem SprachGIS des REDE-Projekts[94] als interaktive Elemente genutzt werden und erlauben so eine dynamische Kombination mit anderen Karten oder Datenquellen, wie zum Beispiel Audioaufnahmen oder Literatur zu einem Ort. Für die Darstellung der Symbole wurde eine spezielle Symbolschrift entwickelt, so dass die Symbole direkt für die Onlineanwendungen mitgeliefert werden. Die digitalisierten Kartendaten sind in der Datenbank in mehrere Tabellen aufgeteilt. In einer Haupttabelle für die Karten werden für jede Karte neben einer eindeutigen ID auch sämtliche relevante Metadaten wie zum Beispiel Kartentitel, Kartennummer und Band gespeichert. Für den MRhSA sind dort 1082 Karten hinterlegt[95]. Die einzelnen Legendeneinträge, die zur Symbolisierung und als Repräsentanten eines linguistischen Phänomens dienen, finden sich in einer Tabelle, die mit der Haupttabelle verknüpft ist. Dort sind 10442 Legendeneinträge zum MRhSA gespeichert. Ein Legendeneintrag besteht aus dem Symbol und einer Legendeneintragsklasse, in der festgelegt ist, ob es sich um ein Hauptsymbol, ein Zusatzsymbol, eine Bemerkung oder eine Abschnittsüberschrift handelt. Weitere Spalten definieren die Größe, Farbe und Anzeigereihenfolge der Einträge oder beinhalten Metainformationen, wie das Erstellungsdatum oder das Datum der letzten Bearbeitung. Die Beschreibung der Einträge ist in einer weiteren Tabelle mit den Legendeneinträgen verknüpft. Über eine *Mapping*-Tabelle werden die Legendeneinträge mit den entsprechenden Orten in der REDE-Datenbank verknüpft. So kann jedem Ort (in einer Karte) das richtige Symbol mit Beschreibung zugeordnet werden. Eine SQL Anfrage erzeugt aus diesen Datensätzen ein GeoJSON Objekt[96] (vgl. Butler u. a. 2016), das durch gängige Geospatialan-

91 <https://www.regionalsprache.de>, abgerufen 23.0.3.2018.

92 Genauer gesagt im Vorläuferprojekt DIWA (<http://diwa.info>, abgerufen 20.05.2018), welches in REDE aufgegangen ist.

93 HPGL (Hewlett-Packard Graphics Language) ist eine Instruktionssprache für die Steuerung von Stiftplottern.

94 Beispielhaft die Karte „weh“ aus Datenserie 1:<https://www.regionalsprache.de/SprachGis/VectorMap/mrhsa/2/133>, abgerufen 18.01.2018.

95 Das REDE System stellt noch weitere Sprachatlanten online zur Verfügung. Mehr Informationen finden sich unter <https://www.regionalsprache.de/atlanten-und-karten.aspx>, abgerufen 18.01.2018.

96 <http://geojson.org>, abgerufen 18.01.2018.

wendungen oder -bibliotheken[97] zur Darstellung von annotierten spatialen Daten (Karten) genutzt werden kann. Die Verfügbarkeit als Onlinepublikation innerhalb eines linguistischen Geoinformationssystems bietet besonders für die „real-time" Analyse einen Mehrwert. Da dieses System als zentrale Sammlung für viele linguistische Karten dient und die Kombination dieser Karten auf einer virtuellen Oberfläche die Analyse von Karten aus verschiedenen Zeitabschnitten deutlich vereinfacht, lässt sich mittels der ebenfalls in diesem System verfügbaren Wenkererhebung eine „real-time"-Analyse über 100 Jahre durchführen (vgl. Herrgen 2010 ).

Da die Datenanalyse dieser Arbeit ontologiebasiert ist, werden die Daten des MRhSA aus der relationalen Datenbank in einen TripleStore[98] überführt. Diese Konvertierung geschieht automatisch mittels eines Transformationsskripts. Dabei werden die erwähnten Tabellen aus der Datenbank ausgelesen und in ein RDF-Datenset umgewandelt, welches wiederum in dem TripleStore gespeichert wird.

Das Skript führt auch eine Reihe von Transformationen durch, da RDF eine flexiblere Strukturierung erlaubt als eine relationale Datenbank. Anstelle über eine ID, die als Markierung für eine Zeile in der Tabelle dient, wird eine URI als Repräsentant für eine *Ressource* generiert. Der gewählte Namensraum für das Datenset ist *<http://issg.de>*. Als *Ressource* werden die Daten aus der REDE-Datenbank aufgefasst, die als Individuum einer ontologischen Klasse dienen können. Einer *Ressource* werden über Statements die zugehörigen Daten in Form von Literalen oder weiteren *Ressourcen* zugeordnet. Anders als in der REDE-Datenbank, wo die Dateneinträge zu einem Ort eine Kombination aus dem Legendeneintrag, der Legendenbeschreibung und dem Ort selbst sind, werden zu jedem Ort die beobachteten Phänomene (im weiteren auch als Observationen bezeichnet) explizit erzeugt. So bekommt jede Observation eine eindeutige URI und lässt sich über diese direkt als *Ressource* referenzieren. Diese Observationen sind einem Ort und einer Karte zugeordnet, die wiederum als *Ressourcen* definiert sind. Eine Legende zu einer Karte wird auch zu einer expliziten *Ressource* und die einzelnen Legendeneinträge, die auch als referenzierbare *Ressource* definiert werden, sind über Relationen mit der Legende und den zugehörigen Observationen verbunden. Mithilfe einer Schemaontologie werden den Daten angemessene ontologische Klassen[99] zugeordnet und in eine einfache Struktur überführt. So wird zum Beispiel definiert, das ein Atlas aus Band oder Karte besteht und ein Band aus Karte. Durch Inferenz werden zudem die Orte, für die Observationen vorliegen, einem Atlas oder Band zugeordnet[100]. Das bedeutet, dass die Orte des Atlas direkt aus den Daten generiert werden und sich bei einer Änderung automatisch anpassen. Die individuellen Karten des MRhSA sind entweder Instanzen der Klassen Lautkarte oder Morpholo-

97 Zum Beispiel: <https://openlayers.org>, abgerufen 23.03.2018, <https://leafletjs.com>, abgerufen 23.03.2018, <https://qgis.org/de/site/>, abgerufen 23.03.2018.

98 Für diese Arbeit wird der TripleStore GraphDB (<https://www.ontotext.com>, abgerufen 18.01.2018.) in der FREE Version verwendet.

99 Im Folgenden mit Kapitälchen markiert. Unmarkierte Bezeichnungen beziehen sich auf Instanzen dieser Klassen.

100 Wenn eine Observation an einem Ort einer Karte zugeordnet ist und die Karte Teil eines Atlas (oder Bands) ist, dann wird dieser Ort von dem Atlas behandelt.

GISCHE KARTE, wobei beide eine Unterklasse von KARTE selbst sind. Alle Observationen selbst sind Instanzen einer LINGUISTISCHEN OBSERVATION, die wiederum eine Unterklasse einer OBSERVATION ist. Die OBSERVATION gehört zu der DataCube-Ontologie[101] und ist ein W3C Vorschlag für die Strukturierung von statistischen Daten. Die informationstragenden Felder in der Datenbank werden über Relationen als LITERAL den Hauptressourcen zugeordnet. Für viele Relationen werden verbreitete oder „best practice"-Vokabulare verwendet. So wird der Dublin Core[102] verwendet zur Beschreibung der Metadaten, wie Kartentitel (*dct:title*) und GeoSPARQL[103], für die Modellierung der Orte und Sprachräume (*geo:Feature*). Dadurch soll eine Schnittstelle[104] zum Semantischen Web geschaffen werden, was eine einfache Verknüpfung mit anderen Wissenssystemen erlaubt. Abbildung 2.1 zeigt einen Ausschnitt aus dem RDF-Datenset zum MRhSA.

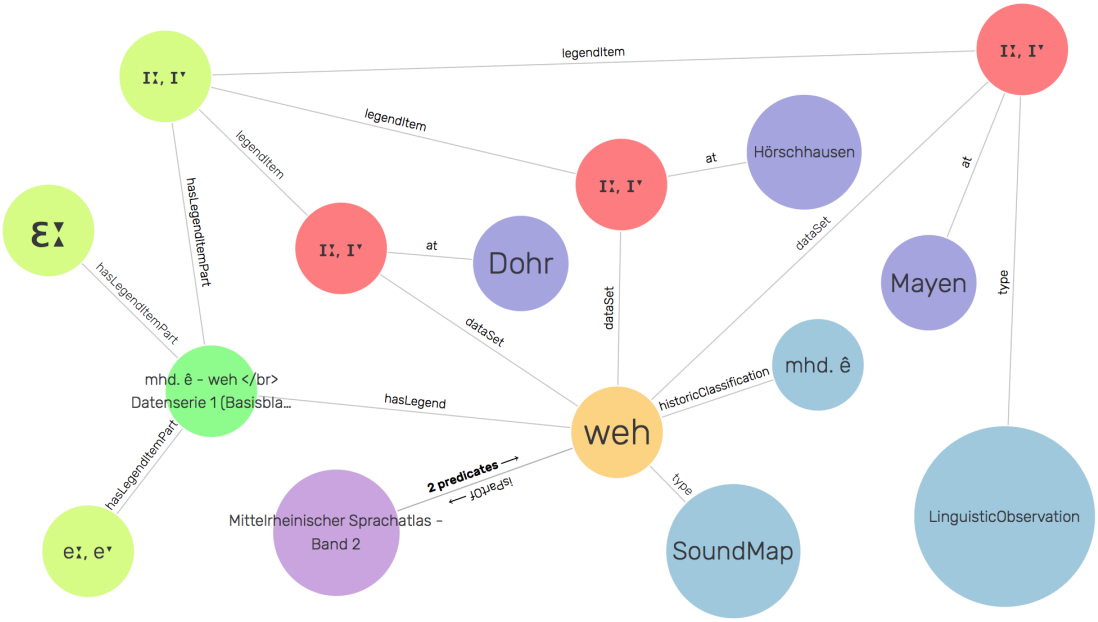



Abbildung 2.1: Ausschnitt aus dem MRhSA-Datenset. Es wird ein Teil der Relationen und Entitäten an den Orten Mayen, Dohr und Hörschhausen, die der Karte „weh" (133) aus Band 2 zugeordnet sind, gezeigt. Erstellt mit dem Visualisierungswerkzeug von GraphDB.

Tabelle 2.1 bietet eine Übersicht über die Anzahl der Instanzen zu den wichtigsten Hauptklassen des Datensets. Die phonetischen Eigenschaften ergeben sich durch Inferenz mittels der *phonOntology*.

101 <https://www.w3.org/TR/vocab-data-cube>, abgerufen 18.01.2018.
102 <http://dublincore.org>, abgerufen 18.01.2018.
103 <http://www.opengeospatial.org/standards/geosparql>, abgerufen 18.01.2018.
104 <https://www.w3.org/standards/semanticweb/data>, abgerufen 18.01.2018.

Tabelle 2.1: Übersicht über die Anzahl der relevanten Daten nach Datenserien getrennt.

| | DATENSERIE 1 | DATENSERIE 2 | GESAMT |
|---|---|---|---|
| Karten | 460 | 276 | 736 |
| Observationen | 248748 | 84386 | 333134 |
| Phon. Eigenschaften | 792068 | 274689 | 1066757 |
| Orte | 546 | 297 | 546 |

## 2.4 DAS LAUTSYSTEM DES MRHSA UND DIE ZUORDNUNG ZU DER PHONETISCHEN ONTOLOGIE

Zusätzlich zu der Schemaontologie, die vornehmlich der Strukturierung der Daten gilt, sind noch zwei weitere Ontologien für den MRhSA angelegt. Einmal die sehr einfach aufgebaute Ontologie der historischen Bezugslaute und die Hauptontologie *phonOntology*. Diese erste Ontologie definiert die historischen Bezugslaute, die als Kategorie einer Karte zugeordnet sind, als ontologische Klasse. Diese Klassen entsprechen den Labeln auf der X-Achse in Abbildung 2.2 und sind in Abschnitt A.3 noch einmal, zusammen mit den zugehörigen Bezugsworten, aufgeführt. Die Wahl der Bezugslaute wurde bereits in Abschnitt 2.1 diskutiert. Die Ontologie unterteilt diese Bezugslaute in Klassen. Zum einen in das HISTORISCHE MITTELHOCHDEUTSCH mit KURZVOKALEN, LANGVOKALEN und DIPHTHONGEN als Unterklassen und zum anderen in das HISTORISCHE WESTGERMANISCH. Es findet jedoch keine historische Analyse oder stärkere Vernetzung der Klassen statt. Diese Ontologie ist die Grundlage bei der Filterung der Daten für die Clusteranalyse, da sie eine einfache Aufteilung der Daten nach den historischen Bezugslauten ermöglicht.

Die *phonOntology* ist die zentrale Ontologie für die Datenanalyse. Diese Ontologie ist dafür verantwortlich, dass die einzelnen Observationen in ihre phonetischen Eigenschaften zerlegt werden. Da zunächst die Observationen nicht den entsprechenden phonetischen Klassen in der Ontologie zugeordnet sind, bedarf es eines Mappings. Für dieses Mapping werden alle distinkten Ausprägungen der Observationen aufgelistet und zusammengefasst. Zu diesen Ausprägungen wird ein entsprechender IPA-Laut gesucht und mit der dazugehörigen Klasse in der *phonOntology* verknüpft. Sollte eine Observation noch zusätzliche phonetische Eigenschaften benötigen, werden diese hinzugefügt. Ein Ausschnitt aus der Mappingtabelle ist in Tabelle 2.3 zu sehen. Insgesamt wurden alle Ausprägungen mit mindestens zehn Observationen für ein Mapping in Betracht gezogen.

Das Erstellen dieser Tabelle ist ein manueller Schritt, und die Ausprägungen erlauben nicht immer eine eindeutige Zuordnung. Häufig wurden uneindeutige Einträge oder mehrere Laute zur Beschreibung des erfassten Phänomens verwendet, so dass selbst mit einem normalisierten Vokabular eine eineindeutige Zuordnung nicht immer möglich ist. Es wurde versucht, sinnvolle Entscheidungen zu treffen und diese konsequent umzusetzen. Bei einer Uneindeutigkeit wurde sich für den häufiger vorkommenden oder den ers-

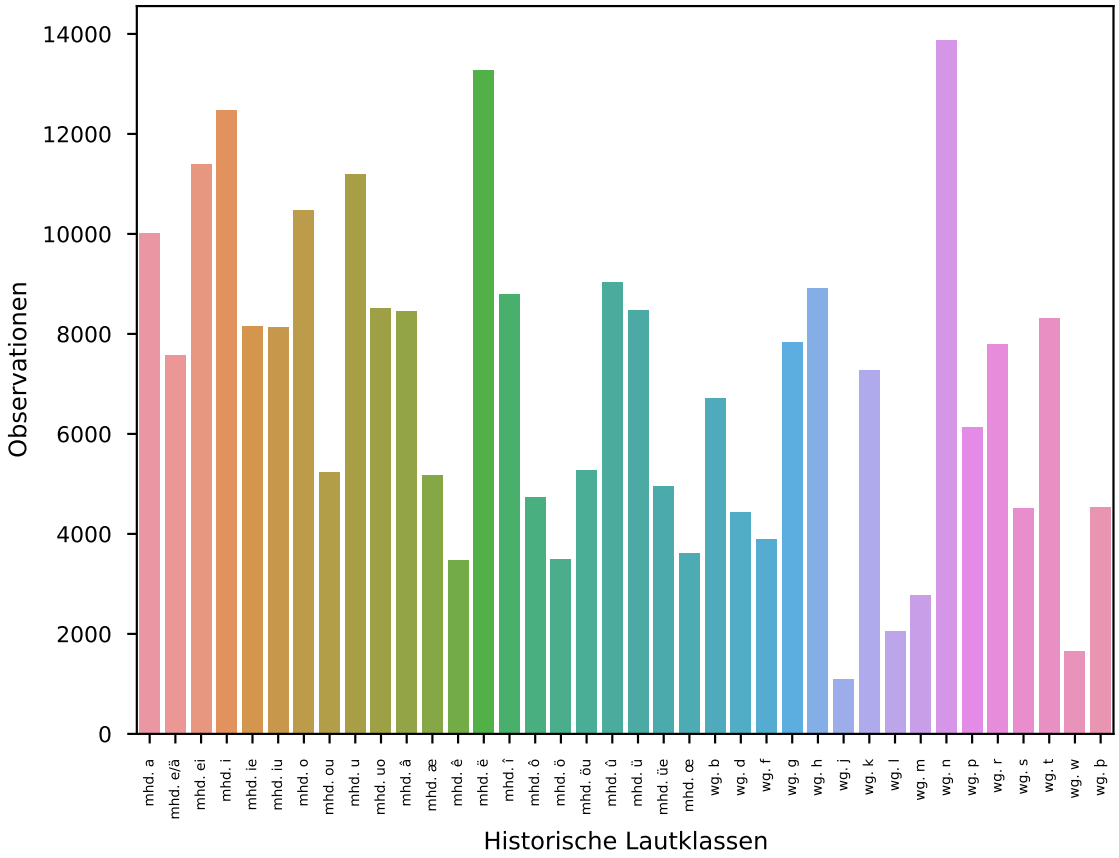



Abbildung 2.2: Verteilung der Observationen aus Datenserie 1 des MRhSA auf die historischen Lautklassen.

ten Laut (bei zwei möglichen) entschieden. So wird zum Beispiel [aː, ɑː, ɑ˙, ɑ̣ː] und Abwandlungen davon in der Ontologie auf das DEUTSCHE A ([aː]) abgebildet. Ausprägungen, zu denen sich kein passender IPA-Laut aus der *phonOntology* finden lies, wurden ignoriert. Insgesamt wurden 464 von 1167 Ausprägungen einem IPA-Laut zugeordnet. Dies entspricht 412645 von 448320 einzelnen Observationen[105] (92% Abdeckung). Eine vollständige Auflistung der Abdeckungen ist in Tabelle 2.2 aufgeführt. Dabei beziehen sich Datenserie 1 und 2 auf die entsprechenden Lautkarten des Datensets und „Alle Observationen“ auf alle Karten.

Dieses Mapping beinhaltet auch die in Abschnitt 2.2 erwähnte Datenglättung. Zum einen wurden nicht alle Observationen erfasst und zum anderen besteht die Möglichkeit, dass Observationen nur unzureichend auf einen IPA-Laut gemappt wurden. Da besonders Ausprägungen, die zur Differenzkartenserie gehören oder nur vereinzelt vorkommen, nicht erfasst wurden, sollte dieser Fehler in Grenzen überschaubar bleiben und kann als Teil des Glättungsprozesses der Clusteranalyse gesehen werden. Die größte Klasse

105 Die Zahlen unterscheiden sich von den in Tabelle 2.1 angegebenen Zahlen, da hier noch die Observationen der Differenzkartenserie einbezogen sind; sofern erkenntlich wurden diese aber beim Mapping ignoriert. Diese Kartenserie trägt für die Analyse keine relevanten Informationen.

von nicht markierten Observationen ist „kein Tonakzent feststellbar“ mit 4908 Einzelausprägungen. Da diese Observationen implizit in den Observationen zu den Tonakzenten enthalten sind, können sie ignoriert werden. Ein unzureichendes Mapping kann durchaus Auswirkungen auf die Ergebnisse haben. Dieser Fehler ist aber nur schwer zu umgehen, da die *phonOntology* nur einen Teil des möglichen Sprachspektrums abbildet und die Observationen nicht immer eine eindeutige Lautmarkierung besitzen. Eine häufige Ausprägung, die bei dem Mapping übergangen wird, obwohl sie mit 2039 Observationen relativ häufig ist, ist die *Heteronymik*, welches die Substitution des Referenzwortes beschreibt. Auch wenn *Heteronymik* durchaus einen Einfluss auf einen Sprachraum haben kann, so lässt sich durch diese Markierung nicht mehr ein Unterschied zu einem Referenzlaut ableiten. Die nächst größere Observationklasse, die ignoriert wird, ist die Zentralisierung, da diese Markierung eine relative Eigenschaft beschreibt, das Mapping sich aber auf absolute Eigenschaften beschränkt. Alle weiteren nicht erfassten Observationen sind Sonder- oder Einzelfälle.

Tabelle 2.2: Übersicht über die Abdeckung der Datenserien mit Observationen, die mit Eigenschaften der *phonOntology* markiert sind.

| | MARKIERTE OBSERVATIONEN | OBSERVATIONEN | ANTEIL |
|---|---|---|---|
| Datenserie 1 | 248733 | 271302 | 92% |
| Datenserie 2 | 84386 | 87081 | 97% |
| Alle Observationen | 412645 | 448320 | 92% |

Diese Mappingtabelle wird mithilfe eines Skriptes eingelesen und jeder Observation wird die entsprechende Klasse aus der *phonOntology* zugeordnet. Die Inferenzengine der Graphdatenbank fügt dann die entsprechenden phonetischen Eigenschaften hinzu. Durch diese Inferenzberechnung werden sehr viele neue Triple der Datenbank hinzugefügt, was die Gesamtmenge der Statements des RDF-Datensets um ungefähr die zehnfache Menge erhöht. Viele dieser hinzugefügten Statements sind trivialer Natur[106]. So wird zum Beispiel zu jeder Observation, die den Laut [ɪː] als Klasse zugeordnet bekommt, zusätzlich inferiert, dass sie zu den Klassen *Vowel* und *Phon* gehören. Auch werden viele anonyme Klassen generiert. So ist [ɪː] äquivalent zu *Phon* mit den Eigenschaften *NearFront*, *LoweredClose*, *Long* und *Unround*. Dieses Konstrukt ist selbst eine anonyme Klasse, und diese Klasse ist zudem eine Unterklasse jeder Teilklasse davon, also unter anderem von *Phon* und *NearFront* und *Phon* und *Long*. Wegen Inferenz ist der Laut damit auch eine Instanz dieser anonymen Klassen. Dies führt dazu, dass eine Observation über 50 anonyme Klassen haben kann. Bei der Analyse spielen diese Klassen aber keine Rolle und werden in der Anfrage, die ein Datenset zur Analyse ge-

106 Aus menschlicher Sicht; für den Computer sind diese gleichwertig zu allen anderen Statements.

neriert, gefiltert. Diese anonymen Klassen werden relevant, wenn Laute nur unter bestimmten Gesichtspunkten analysiert werden sollen. So sind diese Klassen Surrogate für Lautmerkmale wie Vorderzungenvokale oder stimmhafte Konsonanten.

Tabelle 2.3: Auszug aus dem Mapping zwischen Observationen im MRhSA und den Klassen der *phonOntology*.

| # | OBSLABEL | CLABEL | CLASS |
|---|---|---|---|
| 23471 | Ausfall des Konsonanten | - | http://issg.de/ontologies/phonetic#Gap |
| 19063 | eː, eˑ | eː | http://issg.de/ontologies/phonetic#LongCloseMidFrontUnroundedVowel |
| 18041 | ɪ | ɪ | http://issg.de/ontologies/phonetic#LoweredCloseNearFrontUnroundedVowel |
| 14755 | ɪː, ɪˑ | ɪː | http://issg.de/ontologies/phonetic#LongLoweredCloseNearFrontUnroundedVowel |
| 12351 | oː, oˑ | oː | http://issg.de/ontologies/phonetic#LongCloseMidBackRoundedVowel |
| 11192 | ʊ | ʊ | http://issg.de/ontologies/phonetic#LoweredCloseNearBackRoundedVowel |
| 10607 | ɛː, ɛˑ | ɛː | http://issg.de/ontologies/phonetic#LongOpenMidFrontUnroundedVowel |
| 10412 | e | e | http://issg.de/ontologies/phonetic#CloseMidFrontUnroundedVowel |
| 9260 | ʊː, ʊˑ | ʊː | http://issg.de/ontologies/phonetic#LongLoweredCloseNearBackRoundedVowel |
| 8817 | ɛ | ɛ | http://issg.de/ontologies/phonetic#OpenMidFrontUnroundedVowel |
| 8537 | n | n | http://issg.de/ontologies/phonetic#VoicedAlveolarNasal |
| 8035 | ə | ə | http://issg.de/ontologies/phonetic#MidCentralUnroundedVowel |
| 7248 | s | s | http://issg.de/ontologies/phonetic#VoicelessAlveolarFricative |
| 6610 | o | o | http://issg.de/ontologies/phonetic#CloseMidBackRoundedVowel |
| 6606 | f | f | http://issg.de/ontologies/phonetic#VoicelessLabiodentalFricative |
| 5766 | aɪ, a̯ɪ, ɑɪ, ɑɪ | a͡ɪ̯ | http://issg.de/ontologies/phonetic#DiphOpenCentralLoweredCloseNearFront |
| 5471 | x | x | http://issg.de/ontologies/phonetic#VoicelessVelarFricative |

Da das direkte Mapping über eine IPA-Zuordnung erfolgt und das Aufteilen der Laute in ihre phonetischen Eigenschaften durch Inferenz geschieht, ist es möglich, viele Observationen in einem Schritt abzuarbeiten. So zeigt Tabelle 2.3, dass die ersten beiden „Laute“ (Ausfall des Konsonanten und [eː, eˑ]) bereits für über 40000 einzelne Observationen verantwortlich sind. So kann eine ansonsten fast unmögliche Aufgabe auf ein überschaubares Maß reduziert werden. Auch ist es möglich, die Eigenschaften und die Kombination von Eigenschaften nachträglich in der Ontologie zu ändern, ohne das Mapping neu erstellen zu müssen. So können verschiedene Versionen erstellt werden, um Kompatibilität zu anderen System, wie der GOLD-Ontology zu gewährleisten. In dem TripleStore werden Mapping, *phonOntology* und die Daten selbst in gesonderten Graphen gespeichert. Auf diese Weise lässt sich das Mapping leicht austauschen, sollen sich Änderungen als notwendig erweisen, ohne dass die anderen Daten davon betroffen werden. Eine Änderung an diesem Mapping zieht allerdings Reinferierung mit sich, die viel Zeit in Anspruch nimmt. Anfragen werden dann gegen die Vereinigung (Default Graph) dieser Graphen gestellt.

## 2.5 DIE SPRACHRÄUME DES MRHSA

Das Untersuchungsgebiet des MRhSA umfasst den linksrheinischen Teil des Bundeslandes Rheinland-Pfalz und das Bundesland Saarland. Dieser Raum ist auch als der mittlere und südliche Teil des „Rheinischen Fächers“(vgl. Bellmann 1994, S. 9–10) bekannt und wird geographisch im Norden durch die Eifel-Schranke, im Süden durch die Selz-Lauter-Schranke und durch die Hunsrückschranke als mittlere Raumtrennung aufgespannt (vgl.Frings 1957, S. 87). Der nördliche Teil des Untersuchungsgebietes kann nach den Ergeb-

nissen von Lameli und Schmidt (vgl. Lameli 2013; Schmidt 2015) dem Historischen Westdeutsch zugeordnet werden, der südliche Teil dem Mitteldeutschen. Sprachhistorisch werden Sprachräume durch Isoglossen getrennt, also historisch untersuchte, markante Laut- oder Wortschatzgrenzen, die eine deutliche Trennung in zwei Regionen erlauben. Für das Untersuchungsgebiet sind das drei Hauptisoglossen. Im Norden trennt die *Dorp/Dorf*-Isoglosse das Moselfränkische vom Ripuarischen ab. Die Hauptgrenze innerhalb des Untersuchungsgebiets ist die *dat/das*-Isoglosse, die das Moselfränkische vom Rheinfränkischen trennt und auch die Grenze zwischen dem Historischen Westdeutschen und dem Mitteldeutschen ist. Das Rheinfränkische selbst wird im Süden durch die *Pund/Pfund*-Isoglosse vom Alemannischen getrennt. Diese Isoglossen sind räumlich durch die Wenkererhebung gut verortet und lassen sich in den entsprechenden Karten[107] nachvollziehen. Diese Isoglossen haben ihren Ursprung in der zweiten deutschen Lautverschiebung, die als der Ausgangspunkt des Hochdeutschen aus dem Germanischen gilt, womit sie sprachhistorisch bedeutend sind. Bei der Einteilung von Sprachräumen anhand von den als wichtig angesehenen Isoglossen stellte sich heraus, dass diese Isoglossen nicht isoliert im Raum stehen, sondern mit anderen korrelieren. So werden die „harten" Grenzen an den Isoglossen etwas aufgeweicht. Anstelle einer Isoglosse wird ein Isoglossenbündel zur Bildung der Sprachräume einbezogen. Die sukzessiven Laut- oder Wortschatzgrenzen in der Nähe einer Hauptisoglosse bilden somit ein Übergangsgebiet zwischen Dialekträumen[108]. Diese Isoglossenbündel korrelieren häufig mit natürlichen oder politischen Grenzen, so finden sich in der bereits erwähnten Hunsrückschranke Isoglossenbündel, die die Trennung zwischen dem Moselfränkischen und dem Rheinfränkischen beschreiben können.

Wiesinger erweiterte die isoglossenzentrierte Sichtweise um eine strukturelle und erlaubt damit eine *strukturelle Einteilung der deutschen Dialekte* (vgl. Wiesinger 1983). Anstelle nach pivotalen Isoglossen, wird nach strukturellen Gemeinsamkeiten innerhalb gestaffelter Isoglossen für die Bildung eines Sprachraums gesucht, zum Beispiel Qualitätsunterschiede in den Vokalen oder prosodische Distinktionen. Für das Untersuchungsgebiet des MRhSA bildet die Tonakzentgrenze eine scharfe Dialektraumgrenze, da sie ausschließlich in dem Gebiet des Historischen Westdeutsch, im Moselfränkischen vorkommt. Die Tonakzente bilden dort ein signifikantes Merkmal, da sie im Deutschen einzigartig sind (vgl. Schmidt 1986). Diese Strukturgrenzen sind nicht notwendigerweise distinkt, sondern können Übergangsgebiete haben. Dies betrifft vor allem großräumige Dialekteinteilungen.

Sowohl die isoglossenbasierte als auch die strukturbedingte Sprachraumeinteilung führt zu ähnlichen Raumstrukturen. Diese Dialekträume lassen sich auch heute noch „intuitiv", trotz stärkerer Standardisierung der Sprache und deutlich höherer Mobilität, räumlich korrekt verorten. Purschke (2011) zeigt in seiner Dissertation mithilfe eines Perzeptionsexperiments, dass Hö-

107 *Dorp/Dorf*: <https://www.regionalsprache.de/SprachGis/RasterMap/wa/505>, *dat/das*: <https://www.regionalsprache.de/SprachGis/RasterMap/wa/472>, *Pund/Pfund*: <https://www.regionalsprache.de/SprachGis/RasterMap/wa/417>, abgerufen 27.02.2018.

108 Exemplarisch zu sehen bei Martin (1914).

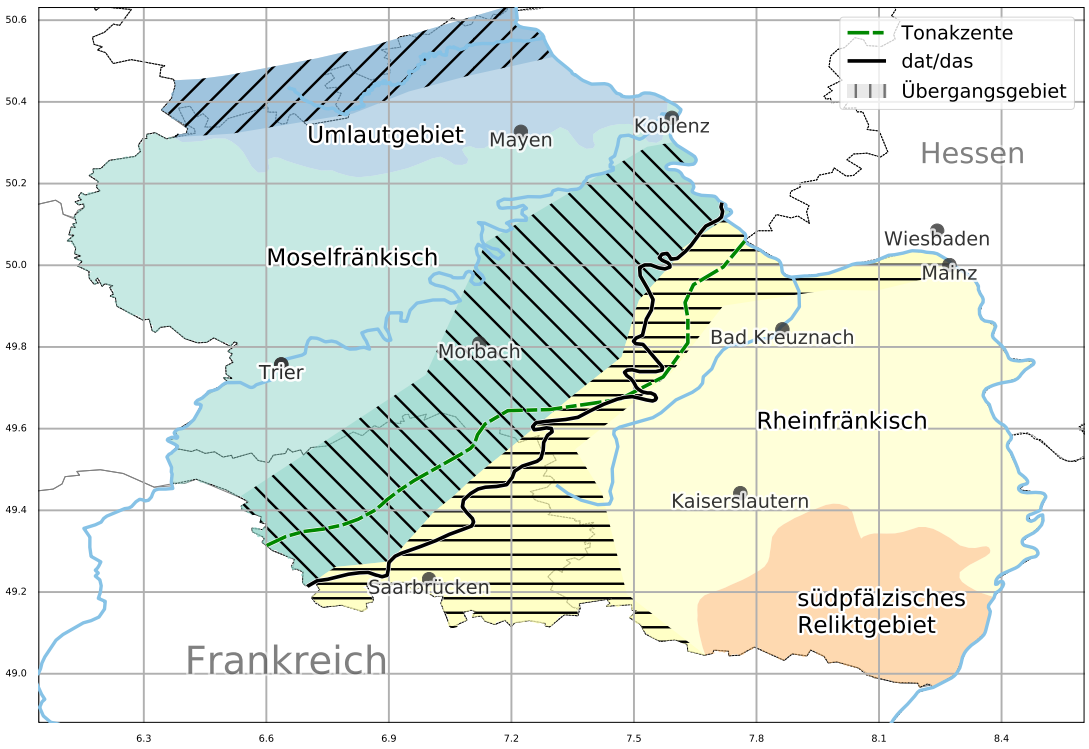



Abbildung 2.3: Das Untersuchungsgebiet des MRhSA mit den Hauptsprachräumen Moselfränkisch und Rheinfränkisch nach der Einteilung von Wiesinger und den zusätzlichen Untersprachräumen Umlautgebiet und südpfälzisches Reliktgebiet nach der Karte *Strukturgrenzen des Westmitteldeutschen* (479b) des fünften Bandes des MRhSA.

rer in der Lage sind, dialektale Sprachproben in die entsprechenden Regionen einzuordnen. Dies trifft besonders auf Hörer in angrenzenden Regionen zu. Da diese Verortung gelingt, obwohl die gewählten Isoglossen und Hervorhebungsmerkmale wie die Tonakzente, nicht allgemein bekannt sind, lässt sich eine tiefere Struktur in dem Dialekt der entsprechenden Region vermuten. Für das Untersuchungsgebiet stellt Purschke zusätzlich fest, dass die beiden Hauptregionen in der Wahrnehmung der lokalen Sprecher als deutlich getrennte und stabile Sprachräume verankert sind (vgl. Purschke 2011, S. 228 ff., 307 ) und über saliente Merkmale, zum Beispiel dem *dat/das*-Gegensatz, referenziert werden können.

Der *Mittelrheinische Sprachatlas* selbst bietet eine gewichtete Einteilung des Untersuchungsgebietes in Dialektregionen basierend auf ausgewählten Karten. Die beiden Hauptregionen sind das bereits erwähnte Moselfränkische im Norden und das Rheinfränkische im Süden. Als Grenze dient dabei die Tonakzentgrenze basierend auf der Karte „Seide“. Diese Grenze folgt ungefähr der *dat/das*-Isoglosse. Die Tonakzente, die im Moselfränkischen markiert mit den Bezeichnern Tonakzent 1 und Tonakzent 2 auftreten, im Rheinfränkischen und in den übrigen deutschen Dialekten allerdings überhaupt nicht, stellen ein deutliches Abgrenzungskriterium dar. Bedingt durch die Tonakzente ergibt sich im Moselfränkischen eine völlig andere

prosodische Grundstruktur, die unter anderem einen Einfluss auf die Realisation von Vokalen hat (vgl. Schmidt 1986). Diese Hauptgrenze wird umrandet von einem Übergangsgebiet, das im Moselfränkischen eine deutlich größere Ausdehnung hat als im Rheinfränkischen. Im Norden wird aus dem Moselfränkischen entlang der sogenannten Entrundungsgrenze, basierend auf der Karte „Füße", ein Untergebiet hervorgehoben. Diese Grenze bildet die zweite Dialektgrenze. Sie trennt das Moselfränkische von dem Ripuarisch-Moselfränkischen-Übergangsgebiet. Dieses nördliche Umlautgebiet zeichnet sich gegenüber dem Moselfränkischen durch konsequent gerundete Vorderzungenvokale aus. Im Süden des Rheinfränkischen hebt sich das südpfälzische Reliktgebiet markiert durch zwei Isoglossen ab. Zum einen zeigen die Isoglossen die Grenze der Phonemverschmelzung, basierend auf der Karte „Frau", und die Grenze der Diphthongierung basierend auf der Karte „froh". Auch wenn diese beiden Phänomene nur wenige Phoneme betreffen, bilden sie doch einen deutlichen Unterschied zum Rest des Rheinfränkischen. Die Phonemverschmelzung markiert den Zusammenfall von Lauten der Lautklassen mhd. ei ~ öu ~ ou zu dem Phonem /ɛː/ innerhalb des Reliktgebietes, während im Rheinfränkischen eine Disktinktion zwischen /ɛː/ und /a̟ː, ɔː/ gewahrt wird. Die Grenze der Diphthongierung basiert auf den Lautklassen mhd. ê ~ œ ~ ô, was eine Distinktion zwischen /eː/ ~ /oː/ im Rheinfränkischen wahrt, wohingegen die Phoneme zu mhd. ô und mhd. â zusammenfallen. In diesem Reliktgebiet haben wir zum einen die Diphthongierung zu /ɛɪ/ ~ /ɔʊ/ und zwischen mhd. ô und mhd. â bleibt ein Phonemunterschied bestehen.

Abbildung 2.4 zeigt das Untersuchungsgebiet des *Mittelrheinischen Sprachatlas* mit den dazugehörigen Isoglossen im Überblick.

Zusätzlich sind noch die *wih/weh*-Isoglosse der Wenkerkarte „weh"[109] und *Korf/Korb*-Isoglosse der Wenkerkarte „Korb"[110] eingezeichnet und die *dat/das*-Isoglosse. Diese Karte dient zudem als Basiskarte für die Clusteranalysen in Kapitel 4. Die Ergebnisse der Clusteranalysen werden mit den in der Karte eingezeichneten Isoglossen in Kontext gesetzt. Die Orte sind nach der Distanz vom geografischen Nordpol aus indiziert, wobei ein o-Index, wie in der Informatik gebräuchlich, angewendet wird.

109 <https://www.regionalsprache.de/SprachGis/RasterMap/wa/113>, abgerufen 07.03.2018.
110 <https://www.regionalsprache.de/SprachGis/RasterMap/wa/289>, abgerufen 07.03.2018.

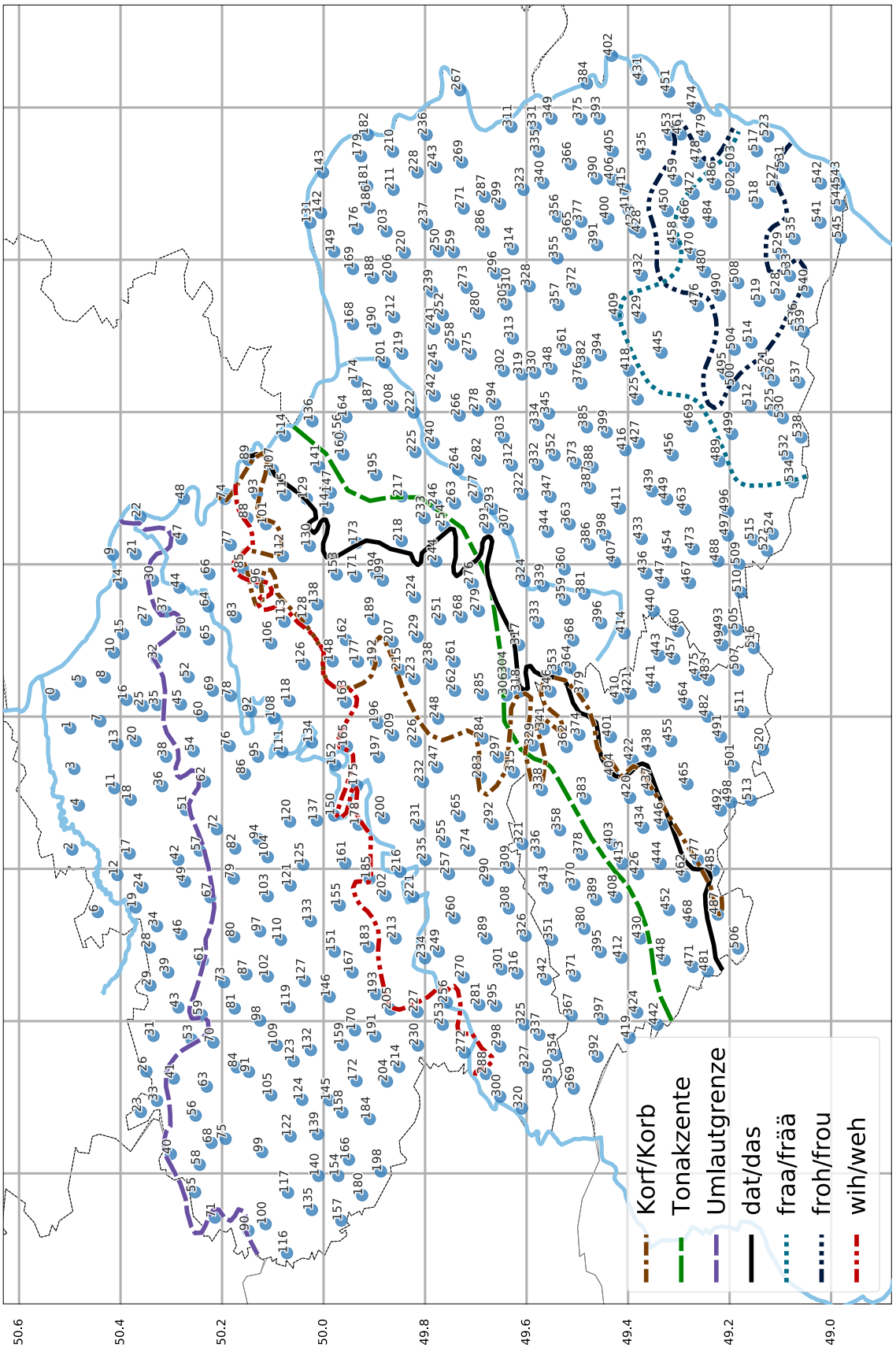



Abbildung 2.4: Das Gebiet des MRhSA mit ausgewählten Sprachgrenzen und den indizierten Orten der Datenerhebung. Die zugehörigen Ortsnamen und die GID für eine Verwendung im REDE SprachGIS finden sich in Abschnitt A.4.

Teil II

# CLUSTERANALYSE

# 3 EINE EINFÜHRUNG IN DIE CLUSTERANALYSE

Dieses Kapitel bietet einen Überblick über ein Teilgebiet des maschinellen Lernens, der Clusteranalyse. Neben einer kurzen Einführung und einer Vorstellung der Hauptmethoden wird ein besonderer Fokus auf die Vorverarbeitung und die Möglichkeiten zur Bewertung der entstehenden Cluster gelegt.

## 3.1 EINLEITUNG

Maschinelles Lernen ist wie Ontologie ein Teilgebiet der künstlichen Intelligenz. Der Fokus liegt hier aber nicht auf dem Entwerfen und Verwalten von Datenstrukturen, sondern auf dem Erkennen von Strukturen in Daten selbst sowie dem Zuordnen von neuen Daten zu bereits annotierten Daten, sogenannten Modellen. Dies geschieht für gewöhnlich mithilfe von Methoden und Prinzipien aus der Mathematik und Statistik. Ontologie hingegen wird eher als Teilgebiet der Logik gesehen. Maschinelles Lernen hat in den vergangenen Jahren stark an Popularität gewonnen, weil neue Entwicklungen im Algorithmendesign (Parallele Algorithmen) und die breite Verfügbarkeit von leistungsfähigen Prozessoren die Verwendung von deutlich komplexeren Algorithmen und Modellen ermöglichen (vgl. Abadi u. a. 2016). Beim maschinellen Lernen wird für gewöhnlich zwischen dem *unüberwachten Lernen* und dem *überwachten Lernen* als den beiden Hauptteilbereichen unterschieden.

Hauptziel des *unüberwachten Lernens* ist es, Strukturen innerhalb einer Datenmenge (Datenset) zu erkennen. Dabei wird nicht auf externes Wissen[111] zurückgegriffen, sondern die angewendeten Methoden versuchen eine Struktur allein auf Basis der verfügbaren Daten zu erkennen. Das Hauptfeld des *unüberwachten Lernens* ist das *Clustering*, das den Kern der Clusteranalyse darstellt. Weitere Gebiete beschäftigen sich mit der Dimensionsreduktion oder -einbettung, die wiederum Vorverarbeitungsschritte für das Clustering sind. Clustering kann als eine Möglichkeit zur Erstellung von sogenannten Modellen gesehen werden, indem es den Daten auf Basis der intrinsischen Struktur Bezeichner[112], meistens in Form von natürlichen Zahlen, zuordnet.

Das *überwachte Lernen* erfordert ein Modell. Dies besteht für gewöhnlich aus einer Datenmenge und dazugehörigen Labels oder Klassen. Ziel ist es, ein System anhand dieses Modells so zu trainieren, dass es Aussagen im Sinne von Klassen- oder Labelzuordnungen über neue Daten machen kann. Dies wird als *Klassifizieren*[113] bezeichnet. Klassifizieren ist bei weitem das größte Feld beim maschinellen Lernen und sehr verbreitet im Bereich der Textklassifikation, Sprach- oder Bilderkennung. Das Erstellen eines Modells

111 Dieses externe Wissen wird auch als Ground Truth bezeichnet.

112 Auch Label oder Klassen genannt.

113 Klassifizieren arbeitet mit diskreten Klassen, bei kontinuierlichen Modellen spricht man von *Regression*.

ist beim Klassifizieren meist der aufwendigste Teil. In Kapitel 5 wird eine Klassifikation der Daten zu der jüngeren Generation im MRhSA vorgenommen. Dabei werden die Ergebnisse der Clusteranalyse aus Kapitel 4 als Modell verwendet.

*Terminologie*

Für ein besseres Verständnis der folgenden Abschnitte werden einige Grundkonzepte und wichtige Terme erklärt. Ein zentraler Term ist das *Datenset*[114], innerhalb der informatischen Auswertung wird es auch als $X$ bezeichnet. Die zu $X$ gehörigen Label oder Klassen werden als $y$ bezeichnet. Diese Bezeichner haben ihren Ursprung in der linearen Algebra, wo Großbuchstaben für Matrizen und Kleinbuchstaben für Vektoren verwendet werden.

$$X = \begin{bmatrix} 1 & \cdots & 3 & 4 \\ 2 & \cdots & 4 & 4 \\ 3 & \cdots & 5 & 3 \end{bmatrix}, y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Eine *Reihe* in $X$, z. B. $\begin{bmatrix} 1 & \cdots & 3 & 4 \end{bmatrix}$, ist ein *Featurevektor*. Ein Featurevektor wird bisweilen auch als *Datenpunkt* bezeichnet, wenn es darum geht, die spatialen[115] Eigenschaften hervorzuheben. So kann zum Beispiel ein Ort im MRhSA durch einen Featurevektor oder Datenpunkt repräsentiert werden. Die Position eines Featurevektors im Datenset markiert mithilfe eines Index einen bestimmten Ort. Die Spalten eines Datensets, z. B. $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$, werden auch *Dimensionen* genannt. Eine Datenzelle, z. B. $\begin{bmatrix} 1 \end{bmatrix}$, wird als *Feature*[116] bezeichnet. Der Klassenvektor $y$ hat dieselbe Länge wie $X$ Reihen hat. In den Zellen dieses Vektors stehen die Labels oder Klassen. Diese werden durch natürliche Zahlen repräsentiert[117]. Die Zahlen gehen für gewöhnlich von 0 bis $k$-1, wobei $k$ die Anzahl der Klassen ist. Mit $k$ wird auch der Parameter bezeichnet, der beim Clustering die Anzahl der zu findenden Cluster angibt. Das Ergebnis eines Clusterings auf einem Datenset $X$ kann also als ein Klassenvektor $y$ betrachtet werden. Häufig basieren Klassenvektoren aber auf der manuellen Zuordnung von Labels. Zum Beispiel können Bildern[118] Label wie „Katze“ oder „Hund“ zugeordnet[119] werden. Basierend auf diesem

114 Der Begriff *Datenset* wurde in Kapitel 1 bereits im Kontext von RDF eingeführt (RDF-Datenset), hier bedeutet Datenset eine geordnete Sammlung von Daten in einer tabellenähnlichen Struktur.

115 Hiermit ist nicht eine geographische Verortung auf einer Karte gemeint, sondern die Position von Vektoren im kartesischen Raum (Vektorraum). Dieser Raum kann durchaus mehr als nur zwei oder drei Dimensionen haben.

116 Auch wenn Feature der gebräuchliche Term in der Fachliteratur ist, werden in dieser Arbeit auch die deutschen Bezeichnungen Eigenschaft oder Merkmal verwendet.

117 Label oder Klassen können zum Beispiel Moselfränkisch und Rheinfränkisch sein. Eine Repräsentation durch Zahlen, beispielsweise (0, 1) wird vorgezogen, da die computerinternen Datenstrukturen beim maschinellen Lernen auf die Speicherung von Zahlen optimiert sind.

118 Ein Bild kann als Featurevektor aufgefasst werden, wobei jedem indizierten Pixel ein Farbwert zugeordnet ist. Dies zeigt außerdem, dass Featurevektoren nicht notwendigerweise einer einfachen Tabellenform folgen müssen, sondern auch komplexere Formen annehmen können. In dem Fall von Bildern setzt sich ein Farbwert aus je drei Unterdimensionen zusammen.

119 Intern wird meistens weiterhin eine Labelzuordung über natürliche Zahlen vorgenommen.

Modell lassen sich Klassifikatoren konstruieren, die in der Lage sind, diese Klassen in Bilder zu erkennen (vgl. Deng u. a. 2009).

## 3.2 DATENVORVERARBEITUNG

Eine Clusteranalyse folgt meistens einem ähnlichen Schema und unterscheidet sich nur in der Art der eingesetzten Clusteralgorithmen. Zunächst müssen die Daten aus einer Datenbank geholt werden. Diese Daten liegen meistens noch nicht in einer quantitativen Matrixform vor, wie es für die meisten Clusteralgorithmen erforderlich ist, sondern bestehen aus den phonetischen Eigenschaften eines Lautes und einer Zuordnung zu einem Ort. Ein sehr kurzer Auszug aus diesen Daten ist in Tabelle 3.1 zu sehen. Die vollständige Tabelle umfasst bis zu 792068 Zeilen für alle phonetischen Eigenschaften zu allen Observationen der Datenserie 1. Die Methodik, um so eine Datensammlung in ein Datenset umzuwandeln, nennt sich *One-Hot-Encoding*. Dabei wird für jede Variable-Ausprägung-Kombination eine eigene Spalte oder Dimension im Datenset angelegt. Für jede dieser Dimensionen wird anstelle der Ausprägung eine 1 gesetzt, alle anderen Dimensionen werden auf 0 gesetzt. So wird die Ausprägung binär codiert[120]. Um einen repräsentativen Vektor (einen Datenpunkt) für einen Ort zu bekommen, addiert man sämtliche so codierten Ausprägungen an einem Ort. Das Ergebnis einer sol-

Tabelle 3.1: Auszug aus den mittels einer SPARQL-Anfrage generierten Rohdaten. Id ist ein Identifikator für einen Ort und p ist die Variable, die phonetische Eigenschaften „bindet“.

| id | p | id | p | id | p | id | p |
|---|---|---|---|---|---|---|---|
| 276 | Long | 184 | Back | 544 | Back | 513 | Round |
| 276 | Back | 184 | Long | 544 | Back | 513 | Back |
| 276 | Back | 184 | NearFront | 544 | Long | 513 | Long |
| 276 | Round | 184 | Long | 544 | Round | 513 | OpenMid |
| 276 | Front | 184 | Long | 544 | Round | 513 | Front |

chen Transformation ist in Tabelle 3.2 zu sehen. Bei einer vollständigen Transformation aller Daten der Datenbankabfrage erhält man einen repräsentativen Featurevektor zu einem Ort, der die Frequenz der einzelnen phonetischen Eigenschaften angibt. Dieses Datenset ist die Grundlage für alle weiteren Vorverarbeitungen und Analysen. Ein so transformiertes Datenset hat für alle 792068 erfassten Lauteigenschaften der Datenserie 1 546 Reihen (Anzahl der Orte) und 47 Dimensionen (Anzahl der Variable-Ausprägung-Kombinationen). Eine Einschränkung nur auf Beobachtungen zum histori-

120 Dies entspricht dem kartesischen Produkt aus Variable und Variablenausprägung. Es ist ersichtlich, dass durch diese Methode sehr viele Dimensionen erzeugt werden können, in denen allerdings die meisten Featurewerte 0 sind. Matrizen, in denen die meisten Werte 0 sind, werden auch als *sparse* (dünnbesetzt) bezeichnet.

schen Langvokalismus für Datenserie 1 erzeugt ein Datenset der Form (546, 23)[121].

Tabelle 3.2: Mittels One-Hot-Encoding transformiertes Datenset basierend auf den Daten aus Tabelle 3.1.

| id | p=Long | p=Back | p=Round | p=Front | p=NearFront | p=OpenMid |
|---|---|---|---|---|---|---|
| 276 | 1 | 2 | 1 | 1 | 0 | 0 |
| 184 | 1 | 3 | 0 | 0 | 1 | 0 |
| 544 | 1 | 2 | 2 | 0 | 0 | 0 |
| 513 | 1 | 1 | 0 | 1 | 0 | 1 |

Der nächste Schritt, der für die Anwendung eines Clusteralgorithmus auf ein Datenset notwendig ist, ist die *Skalierung* der Daten. Dabei wird die Verteilung der Daten auf eine vergleichbare Skala gebracht. Viele Klassifizierungs- und Clusteralgorithmen basieren auf Distanz- oder Wahrscheinlichkeitsbestimmungen in einem Vektorraum. Damit die Distanz zwischen zwei Datenpunkten gemessen werden kann, sollten sich alle Dimensionen in einem vergleichbaren Spektrum befinden. So soll verhindert werden, dass einzelne Dimensionen einen zu starken Bias[122] bieten. Für gewöhnlich transformiert man die Daten so, dass eine angenommene Normalverteilung einen Mittelwert und Median von 0 sowie ein oberes und unteres Quartil von 1 respektive -1 hat. Für das Datenset bedeutet das, dass auf jede Spalte folgende Formel angewendet wird:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

wobei $x^{(i)}$ die i-te Dimensionspalte, $\mu_x$ das arithmetische Mittel der Spalte und $\sigma_x$ die entsprechende Standardabweichung ist. Bei realen Daten können Mittelwert und Median durchaus verschieden sein und die Quartilen müssen nicht notwendigerweise bei 1 und -1 liegen. Die Datenverteilung eines Datensets kann als Boxplot dargestellt werden. Abbildung 3.1 zeigt eine derartige Datenverteilung. Man erkennt deutlich die vielen Ausreißer.

Ausreißer können für Clusteralgorithmen ein Problem darstellen, da sie den Entscheidungshorizont für ein Cluster beeinflussen können, andererseits tragen diese Ausreißer aber auch linguistische Informationen und dürfen nicht einfach ignoriert werden. Für die vorliegende Datenanalyse wurde als Kompromiss eine Kappungsfunktion für Ausreißer eingeführt. So werden alle Werte, die größer als der Wert der oberen Quartile plus der dreifachen Interquartilenreichweite sind, auf diesen Wert gesetzt und Werte, die kleiner als die untere Quartile minus der dreifachen Interquartilenreichweite sind, ebenfalls. Dadurch bleiben Ausreißer noch bestehen, insgesamt wird

121 Diese Schreibweise ist eine gebräuchliche Art, die „Form“ eines Datensets zu beschreiben. Dabei gibt das erste Element des Tuples die Anzahl der Datenpunkte und das zweite die Anzahl der Dimensionen an.

122 Es ist bei den meisten realen Datensets nicht möglich, den Bias völlig zu eliminieren, besonders wenn manche Dimensionen nur über ein kleines Spektrum verfügen. Die beste Möglichkeit, gegen Bias anzugehen, ist eine größere Datenmenge..

der Datenraum aber normalisiert, was beim Clustering helfen kann[123]. Ein so transformiertes Datenset kann nun von den meisten Verfahren des maschinellen Lernens verarbeitet werden.

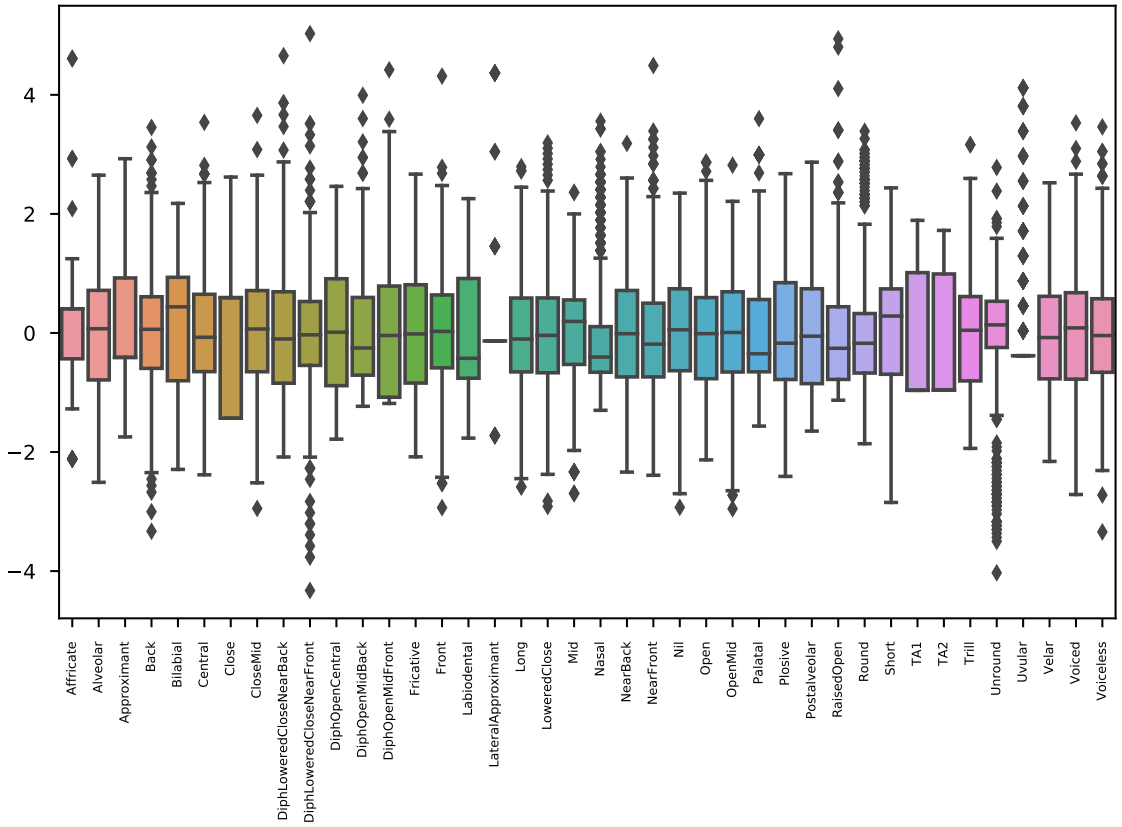



Abbildung 3.1: Verteilung der einzelnen Phänomene für alle Observationen der Datenserie 1. Der Mittelwert ist bei 0. Der Strich innerhalb der Boxen markiert den Median (Q2), die Ausdehnung der Box die 0.25 (Q1) bzw. 0.75 (Q3) Quartile. Die Länge der dünnen Striche beschreibt die Reichweite der 1.5-fachen Interquartilenreichweite (IQR) ausgehend von Q1 bzw. Q3. Werte außerhalb dieses Bereichs werden als Ausreißer bezeichnet.

Häufig wird noch ein weiterer Schritt eingefügt, eine sogenannte Dimensionsreduktion. Eine hohe Anzahl von Dimensionen kann gewisse Probleme mit sich bringen. Das Hauptproblem ist unter dem Schlagwort *Fluch der Dimensionalität* (vgl. Nasrabadi 2007, S. 33) bekannt und beschreibt den Umstand, dass bei vielen Dimensionen die Chance stark ansteigt, eine große Entfernung in einer Richtung zwischen zwei Datenpunkten zu haben, da zwei Werte innerhalb einer Dimension sehr weit auseinander liegen können. Dies wiederum bereitet Probleme beim Erstellen von Clustern, die auf distanzbasierten Algorithmen basieren. Eine Dimensionsreduktion kann zu-

123 Vergleiche zwischen gekappten und ungekappten Datensets haben keine nennenswerten Unterschiede im Raum gezeigt. Sowohl T-Tests (vgl. Lowry 2014a) auf identische Mittelwerte als auch ein Levene-Test (vgl. Brown und Forsythe 1974) auf Grundmengen mit gleichen Varianzen liefern hohe p-Werte. Dies stützt die 0-Hypothese, dass die Datensets auf derselben Verteilung beruhen. Gekappte Werte liefern stabilere Cluster.

dem redundante Dimensionen zusammenfassen oder triviale weglassen[124]. Dies führt zu einer *Glättung* des Datensets, dies kann helfen, stabilere und besser trennbare Cluster zu finden. Die bekannteste Dimensionsreduktionstechnik ist die Hauptkomponentenanalyse[125] (PCA) (Jolliffe, 1986; Tipping und Bishop, 1999). Die PCA ist eine Hauptachsentransformationstechnik, bei der eine neue $n \leq N$-dimensionale Vektorraumbasis so erzeugt wird, dass die Varianz des originalen N-dimensionalen Datensets möglichst gut durch Vektoren in dem neuen n-dimensionalen Vektorraum abgebildet wird. Abbildung 3.2 zeigt, wie viel Varianz des ursprünglichen Datensets durch ein PCA-transformiertes Datenset erklärt wird. Dieses Beispiel demonstriert, dass eine Reduktion auf 28 Dimensionen 99% der Varianz des ursprünglich 47-dimensionalen Datensets erklärt und die ersten drei Dimensionen bereits 60%. Die neuen Dimensionen sind allerdings nicht mehr eindeutig phonetischen Eigenschaften zugeordnet, sondern sind eine gewichtete Kombination aller Phänomene. Der Einfluss einer alten Dimension auf die neuen lässt sich allerdings berechnen. Dieser Umstand kann bei der Gewichtung der einzelnen ursprünglichen Dimensionen hilfreich sein. Ein hoher Einfluss auf die neuen Dimensionen bedeutet, dass das phonetische Phänomen wichtig ist. Zudem sind die neuen Dimensionen selbst nach Einfluss geordnet, ein hoher Einfluss auf eine niedrig indizierte Dimension ist wichtiger als der Einfluss auf eine höher indizierte.

Die Hauptkomponentenanalyse wird als Vorverarbeitungsschritt für die Clusteranalyse des MRhSA angewendet. Dabei ist der Algorithmus so eingestellt, dass die neuen Dimensionen 99% der Varianz erklären sollen. Das dadurch entstehende Datenset dient als das transformierte Datenset, welches für die Clusteranalyse verwendet wird. Weiterführende Analysen verwenden teilweise das nicht transformierte Datenset, insbesondere wenn der Einfluss von Features auf ein Clustering betrachtet wird.

## 3.3 CLUSTERING

Clustering beschreibt den Versuch, mittels bestimmter mathematischer oder statistischer Prinzipien Strukturen in einem Datenset zu finden. Die Anzahl der zu findenden Strukturen (Cluster) wird meistens als Kenngröße $k$ dem Clusteralgorithmus vorgegeben. Da Clustering zu den unüberwachten Lernmethoden zählt, ist häufig über die optimale Wahl des Parameter $k$ wenig bekannt. Auch ist es schwierig, bei Clustering von „richtig“ oder „falsch“ zu sprechen, da die Algorithmen k Cluster basierend auf den zugrundeliegenden Prinzipien finden werden. Ohne eine *Ground Truth*, wie zum Beispiel historische Quellen, muss man sich auf interne Maße zur Wahl und Bewertung der Anzahl der Cluster verlassen. Dies wird in Abschnitt 3.4 genauer besprochen. Clustering lässt sich grob in drei Gruppen unterteilen, wobei die dritte Gruppe in dieser Arbeit nicht weiter zum Einsatz kommt und nur kurz erwähnt wird. Die Hauptgruppen sind das *partitionierende Clustering* und das *hierarchische Clustering* (vgl. Jain 2010). Beim partitionierenden Cluste-

124 Eine Dimension, in der alle Features 1 sind, trägt keine relevante Information für ein Clustering.

125 Engl.: Principal Component Analysis.

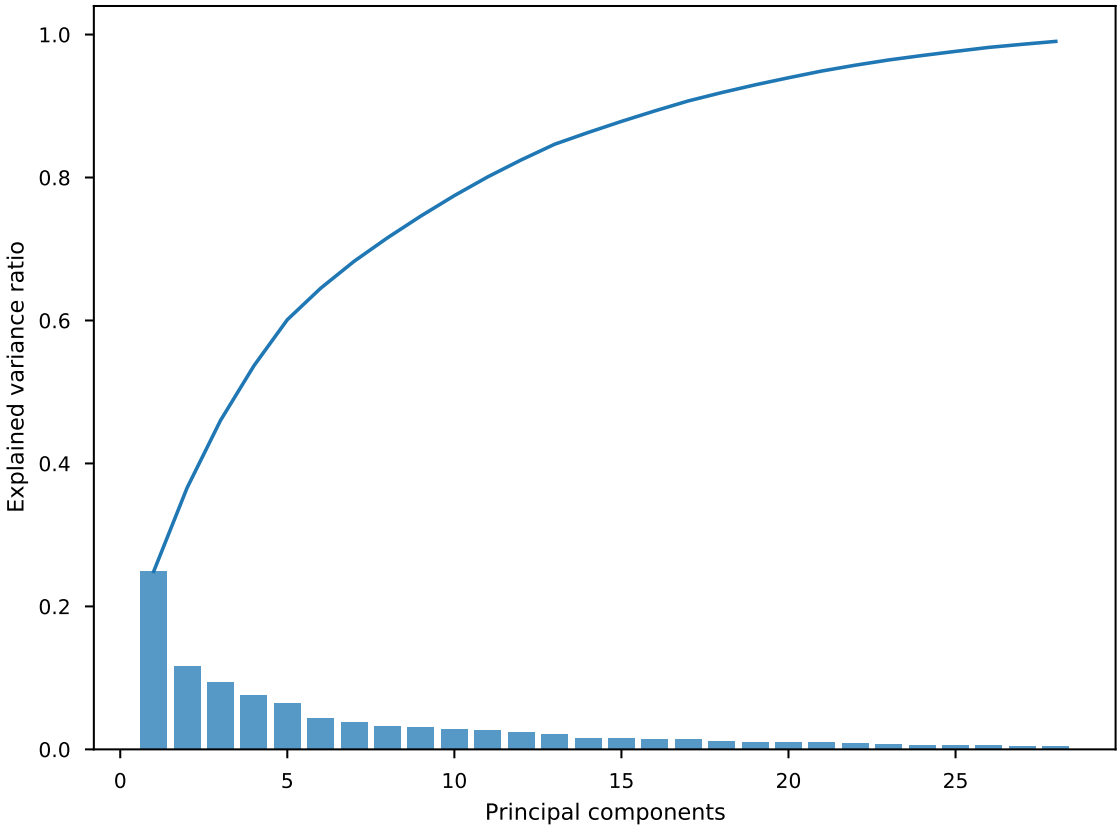



Abbildung 3.2: Anteile der neuen Dimensionen an der Gesamtvarianz des Datensets über alle phonetischen Eigenschaften der Datenserie 1 nach einer PCA.

ring werden Cluster erzeugt, indem der Raum, in dem das Datenset eingebettet ist, in Teilräume unterteilt wird. Dies geschieht meistens mithilfe von Distanz- oder Wahrscheinlichkeitsmaßen, die in iterativen Prozessen optimiert werden. Beim hierarchischen oder agglomerativen Clustering werden die Cluster entweder durch das rekursive Teilen des Datensets anhand einer *Metrik* (hierarchisches Clustering) oder dem sukzessiven Zusammenführen der einzelnen Datenpunkte mittels einer Metrik (agglomeratives Clustering) erzeugt, bis die gewählte Clusteranzahl erreicht ist. Das vollständige Zerlegen respektive Zusammenführen eines Datensets erzeugt ein *Dendrogramm*. Das Zerlegen eines Datensets wird auch *Top-Down*-Ansatz genannt, das zusammenführen *Bottom-Up*. In der Praxis dominiert das agglomerative Clustering, da es etwas effizienter implementierbar ist.

Die dritte Gruppe ist *dichtebasiertes Clustering*. Ein Merkmal dieser Gruppe ist, dass es ohne einen Parameter $k$ auskommt, stattdessen wird die Anzahl der Cluster über eine *Dichtefunktion* bestimmt. Abhängig von der Dichtefunktion werden mehr oder weniger Cluster ermittelt, außerdem ermöglicht es das Erkennen von Ausreißern. Da die Dichtefunktion aber sehr stark von der Form des Datensets abhängt, ist das Finden einer optimalen Anzahl von Clustern häufig deutlich komplexer als mit den anderen beiden Clusteringansätzen. Zusätzlich zu diesen drei Hauptgruppen existieren noch zahl-

reiche Hybrid- oder Kombinationsmethoden, die versuchen, die jeweiligen Schwächen der einzelnen Gruppen zu minimieren.

Die in dieser Arbeit verwendeten Clusteralgorithmen werden im Folgenden vorgestellt.

*K-Means*

K-Means ist der bekannteste Clusteralgorithmus. Erstmals namentlich erwähnt wurde er von MacQueen (1967). Die Prinzipien wurden aber bereits zehn Jahre früher von Steinhaus (1956) und Llyod[126] beschrieben. Die Grundannahme basiert auf der Existenz von $k$ Clusterzentren. Zu den Datenpunkten eines Datensets wird ein sogenannter *Fehler*[127] in Form des Quadrats der Abstände zu den Clusterzentren ermittelt. Ziel ist es, die Clusterzentren so zu finden, dass die Summe aller dieser Fehler minimal ist. Diese Methode wird von Lloyd (1982) als *Least-Square-Method* beschrieben. Der generelle Ablauf des Algorithmus ist einfach. Zunächst werden zufällig im Datenraum des Datensets $k$ Clusterzentren verteilt. Nun wird der Fehler der Datenpunkte zu den Clusterzentren ermittelt. Jeder Datenpunkt wird dem Clusterzentrum zugeordnet, bei dem sein Fehler am geringsten ist. Die Summe über alle Fehler wird auch als *Inertia* oder *Sum of Square Means* (SSE) bezeichnet und berechnet sich wie folgt:

$$\text{Inertia} := \sum_{i=0}^{n} \min_{\mu_j \in C} (\|x_j - \mu_j\|^2)$$

wobei gilt, dass $C$ eine Menge von Clustern und $\mu_j$ ein Clusterzentrum ist. Mit $\|x_j - \mu_j\|$ wird eine Distanzmetrik denotiert, in diesem Fall die Differenz zwischen dem Datenpunkt und dem Clusterzentrum in Form der euklidischen Distanz. Ziel des Algorithmus ist es, diese Summe zu minimieren. Nach der initialen Berechnung dieser SSE werden die Clusterzentren angepasst. Dazu wird für alle Cluster die jeweilige konvexe Hülle der Datenpunkte, die dem aktuellen Clusterzentrum zugeordnet sind, erzeugt und das Zentrum dieser Hülle als neues Clusterzentrum gesetzt. Dann erfolgt eine neue Zuordnung und Fehlerberechnung auf Basis dieser neuen Clusterzentren. Dies wird wiederholt, bis es zu keiner Änderung mehr kommt. Dieser Ablauf zeigt auch ein paar Schwächen des K-Means-Algorithmus auf. Zum einen sollte das Datenset *konvex*[128] sein und zum anderen kann diese Art der Optimierung zu verschiedenen Ergebnissen abhängig von den initialen Clusterzentren führen. Dies wird für gewöhnlich umgangen, indem man den Algorithmus mit verschiedenen Initialzentren wiederholt und das häufigste Ergebnis als Endergebnis auswählt.

126 Allerdings erst 1982 von Lloyd (1982) veröffentlicht.

127 Beim maschinellen Lernen werden Fehler häufig in Form eines Abstandes von einem Datenpunkt zu einem Zielwert beschrieben.

128 Es ist mit K-Means nicht mögliche, Cluster der Art „inneres"-Cluster–„äußeres"-Cluster zu finden. Ein Problem mit *konvex* ist, dass wir keine räumliche Vorstellung von Dimensionen mehr als drei haben und wir uns deshalb *konvex* für höherdimensionale Räume nicht einmal vorstellen können.

*Aggrolomatives Clustering nach Ward*

Aggrolomatives Clustering (vgl. Jain und Dubes 1988) basiert auf dem sukzessiven Zusammenführen von Clustern zu immer größer werdenden Clustern. Zu Beginn des Algorithmus ist jeder Datenpunkt sein eigener Cluster. Der Algorithmus fasst so lange Cluster nach einer bestimmten Strategie zusammen, bis nur noch die gewünschten $k$ Cluster übrig sind. Das iterative Zusammenfassen der Cluster lässt sich auch als Baumstruktur auffassen, die bei komplettem Durchlauf als Dendrogramm bezeichnet wird. Ein solches Dendrogramm erlaubt eine gewisse Einsicht in die Datenstruktur und kann sehr hilfreich bei der Wahl von $k$ sein oder bei einer Analyse von Ähnlichkeiten zwischen Clustern. Abbildung 3.3 zeigt ein derartiges Dendrogramm. Man kann bereits erkennen, dass es zwei Hauptäste gibt und der zweite Ast (in dieser Darstellung der untere) sich noch einmal in zwei Unteräste teilt. Man kann also vermuten, dass sich für dieses Datenset ein Zweier- oder Dreierclustering anbietet.

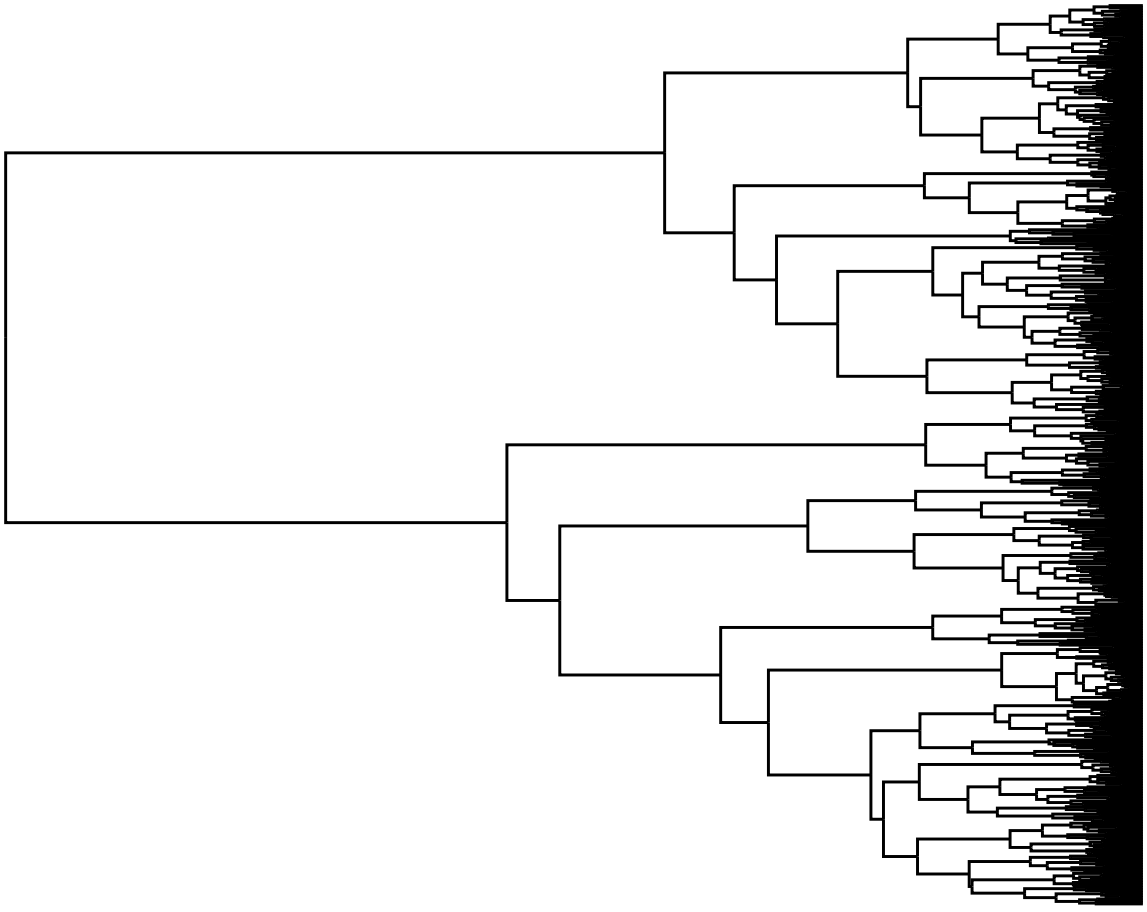


Abbildung 3.3: Dendrogramm über die Datenpunkte für alle phonetischen Eigenschaften. Basierend auf den euklidischen Distanzen und der Ward-Linkage.

Die Strategie, nach der die Datenpunkte vereint werden, wird auch als *Linkage* bezeichnet. Es gibt verschiedene Linkagestrategien, die verbreitetste ist das Linkage nach Ward (vgl. Ward 1963).

SINGLE-LINKAGE Zwei Cluster werden vereinigt, wenn der Abstand zwischen den beiden nächsten Punkten in den jeweiligen Clustern minimal ist.

COMPLETE-LINKAGE Zwei Cluster werden vereinigt, wenn der Abstand zwischen den beiden am weitesten entfernten Punkten in den jeweiligen Clustern minimal ist.

AVERAGE-LINKAGE Zwei Cluster werden vereinigt, wenn der mittlere Abstand aller Punkte aus dem einen Cluster zu den Punkten aus dem anderen Cluster minimal ist.

UPGMA Zwei Cluster werden vereinigt, wenn der mittlere Abstand aller Punkte aus beiden Clustern zusammen minimal ist. Mit anderen Worten, die Cluster, die die flächenmäßig geringste konvexe Hülle aufspannen, werden vereinigt.

WARD-LINKAGE Zwei Cluster werden vereinigt, so dass sich die Gesamtvarianz durch das Zusammenführen dieser Cluster nur minimal erhöht. Die Gesamtvarianz ist dabei wie bei K-Means die SSE mit den jeweiligen Clustermittelpunkten als Clusterzentren. Diese Methode erzeugt größere, aber auch ausgeglichenere Cluster.

In dieser Arbeit wird ausschließlich das Ward-Linkage für das hierarchische Clustering verwendet.

*Gaussian Mixture Models*

Das Gaussian Mixture Model (GMM) (vgl. Reynolds und Rose 1995) basiert auf der Annahme, dass sich innerhalb einer Datenmenge $k$ Untermengen befinden, die sich mithilfe der *Erwartungsmaximierung* hervorheben lassen. Dabei wird zusätzlich angenommen, dass diese Untermengen normalverteilt[129] sind. Erwartungsmaximierung (Expectation–Maximization oder EM-Algorithmus) (vgl. Moon 1996) ist ein Algorithmus, bei dem zunächst, basierend auf zufällig initialisierten Parametern, $k$ Unterverteilungen im Datenset berechnet werden (Erwartung) und anschließend die Wahrscheinlichkeit berechnet wird, dass diese Verteilungen vorliegen. Anhand einer Fehlerfunktion werden die Parameter nun angepasst (Maximierung). Dieser Prozess wird solange wiederholt, bis die Wahrscheinlichkeit für das Vorliegen der Unterverteilungen ein Maximum erreicht hat. Abbildung 3.4 zeigt die Anwendung eines GMM auf das Datenset zu den historischen Langvokalen des Mittelhochdeutschen. Die Ellipsen sind Hilfslinien, die die negative logarithmierte Wahrscheinlichkeit angeben, dass ein Punkt zu einem Cluster gehört. In der Darstellung werden Ellipsen verschmolzen, wenn sie sich schneiden. Man sieht an dieser Abbildung aber auch, dass ein Clustering von „realen" Daten sich von Beispieldatensets[130] unterscheiden kann. Es lassen sich zwar

129 Normalverteilung wird auch als Gauß-Verteilung bezeichnet.
130 Engl.: Toy datasets.

durchaus zwei Hauptcluster in den Daten erkennen, aber die Trennung ist nicht völlig distinkt[131].

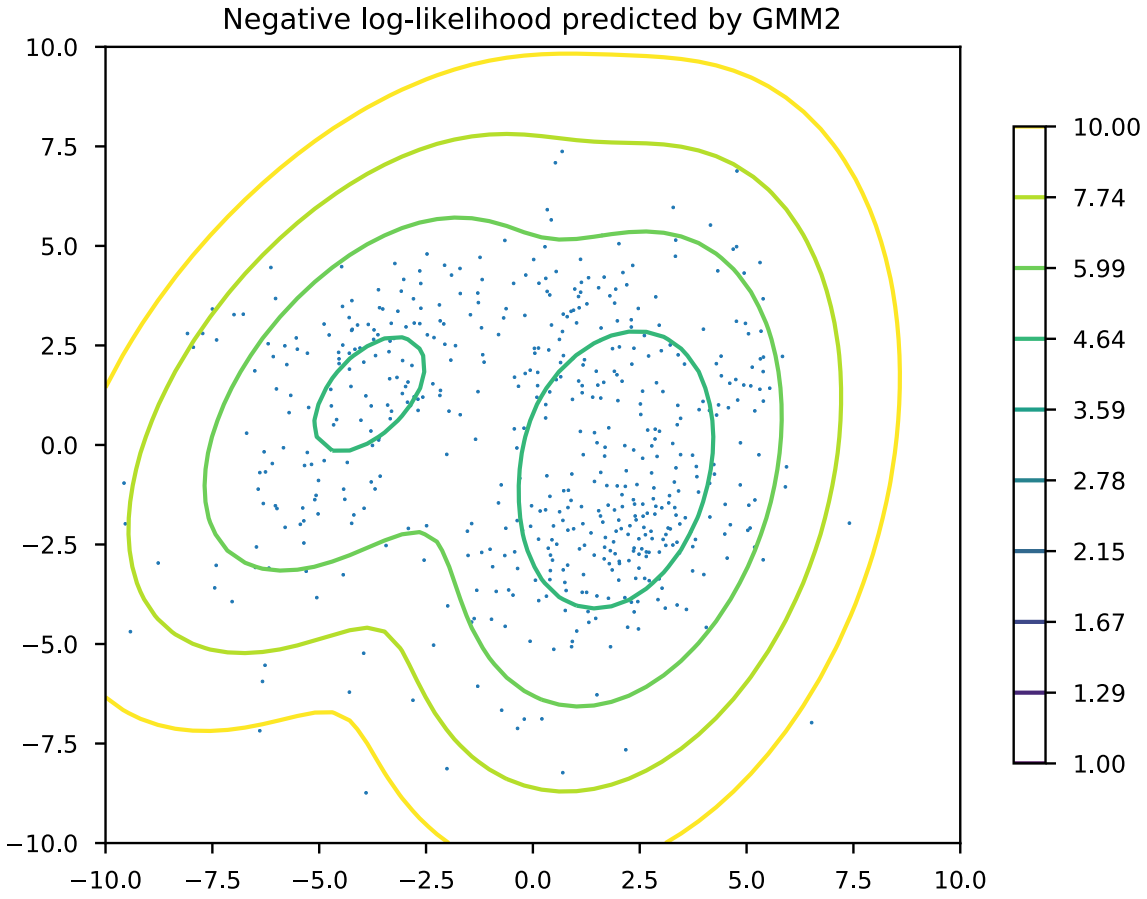



Abbildung 3.4: Beispiel für die Anwendung des Gaussian Mixture Model auf dem Datenset zu den historischen Langvokalen des Mittelhochdeutschen und k=2. Um eine Visualisierung zu ermöglichen wurde das Datenset mittels Multidimensionaler Skalierung (MDS) auf zwei Dimensionen reduziert.

## 3.4 CLUSTERVERIFIKATION

Anders als beim Klassifizieren, bei dem man über ein Referenzmodell verfügt, mit dem die gegebenen Daten verglichen werden können, steht beim Clustering nur ein Datenset zur Verfügung und man muss andere Wege finden, die Ergebnisse zu bewerten. Wie bereits erwähnt kann ein Clustering ein Modell generieren. Allerdings bietet so ein generiertes Modell zunächst keine Hinweise, ob es eine Fragestellung korrekt abbildet und kann bestenfalls als eine Hypothese angesehen werden. Dabei ist zu beachten, dass diese Hypothesen auf mathematischen oder statistischen Strukturen innerhalb

131 Dieses Clustering basiert auf einem modifizierten Datenset, welches durch eine Dimensionsreduktion auf zwei Dimensionen reduziert wurde und es bildet nur eine ungefähre Annäherung an die „reale“ Datenstruktur ab.

des Datensets beruhen. In dieser Hinsicht kann ein Clustering als korrekt angenommen werden, ob sich dies auch in eine Interpretation der Struktur übertragen lässt, ist dadurch aber nicht geklärt. Mit Aussagen über die Struktur des Clusterings lassen sich intrinsische Verifikationsmetriken angeben. Man kann zum Beispiel überprüfen, wie groß die Dispersion[132] eines Clusters ist. Anschaulich bedeutet das, ob sich die Datenpunkte um ein Zentrum bündeln, innerhalb eines Clusters verteilen oder eher in der Grenzregion eines Clusters liegen. Eine weitere Möglichkeit ist die Überprüfung der Stabilität eines Clusters. Ein stabiles Clustering erzeugt ähnliche Cluster, wenn nur eine zufällig ausgewählte Teilmenge der ursprünglichen Daten zur Verfügung steht. Mit Methoden wie Bootstrapping oder Crossvalidation, also dem wiederholten Anwenden der Algorithmen auf zufällige Teilmengen des Datensets, lassen sich Aussagen zur allgemeinen Stabilität treffen. Für extrinsische Vergleiche bieten sich historische Quellen oder zur Zeit gängige Annahmen an. So besteht eine wichtiger Teil der Clusteranalyse zum MRhSA auf dem Vergleich mit den derzeit in der Dialektologie gesetzten Raumstrukturen. Diese externen Quellen können als eine Art *Ground Truth* agieren. Dabei ist zu beachten, dass statistische Datenanalysen häufig angewendet werden, um genau die Behauptungen der externen, empirischen Quellen zu überprüfen, weshalb diese Quellen nicht als *Ground Truth* agieren können, sondern nur als Vergleichsmodell. Zusätzlich zu diesen Strukturuntersuchungen können noch sogenannte Raumeinbettungen[133] erstellt werden. Embeddings sind auch Dimensionsreduktionstechniken wie die PCA und dienen dazu, die Datenpunkte in einem zweidimenionalen Raum, wie zum Beispiel einem Streuungsplot, abzubilden. Dabei wird besonderer Wert darauf gelegt, dass die daraus resultierende räumliche Struktur eine möglichst gute Repräsentation der hochdimensionalen Daten ist. Dies kann natürlich nur eine Annäherung sein. So sieht man in Abbildung 3.2, dass zwei Dimensionen ungefähr 50% der Varianz erklären und obwohl das bereits eine Menge für zwei Dimensionen (von den ursprünglich 47) ist, heißt das umgekehrt auch, dass eine solche Darstellung nur die Hälfte der tatsächlichen Datenstruktur widerspiegelt. Zur Verifikation von Clustern sollten mehrere Methoden herangezogen werden, damit man sich ein besseres Gesamtbild schaffen kann.

*Silhouettenkoeffizient*

Der Silhouettenkoeffizient ist ein Maß für die mittlere Clusterdispersion. Die Methode wurde erstmalig von Rousseeuw (1987) beschrieben und dient als Metrik, um Aussagen über einzelne Datenpunkte in einem Clustering oder über das komplette geclusterte Datenset zu machen. Eine Silhouette für einen Punkt $i$ wird berechnet, indem man die mittlere Distanz $a$ zu allen anderen Punkten des Clusters mit der niedrigsten mittleren Distanz $b$ des Punktes zu Punkten aus anderen Clustern vergleicht.

$$\text{Silhouette} = (b - a) / \max(a, b)$$

132 Auch Streuung oder Varianz genannt.
133 Engl.: Embeddings.

Das Ergebnis ist ein Wert zwischen -1 und 1, wobei ein negativer Wert auf eine mögliche Falschzuordnung des Punktes schließen lässt. Hohe Werte bedeuten eine gute Nähe zum Clusterzentrum und Werte nahe 0 eine Nähe zum Rand des Clusterings. Der Silhouettenkoeffizient wird aus dem Durchschnitt aller Silhouetten gebildet. Abbildung 3.5 zeigt eine Visualisierung der einzelnen Silhouetten und des Silhouettenkoeffizienten als gestrichelte rote Linie. Es ist zu beachten, dass in der Literatur für gewöhnlich ein Silhouettenkoeffizient>0.5 als ein gutes Clustering angegeben wird. Die Werte dieser Clusteranalyse liegen deutlich darunter (in Bereichen um 0.2). Man muss jedoch berücksichtigen, dass es sich bei den verwendeten Sprachdaten um keine völlig isolierten Gebilde handelt, sondern Berührungsgebiete, in denen Sprachkontakt stattfindet, zu erwarten sind. Das Konzept der Übergangsgebiete wurde in Abschnitt 2.5 kurz angesprochen. Deswegen sollten die durch das Clustering bestimmten Grenzen eher als Grenzregionen und nicht als distinkte Grenzen aufgefasst werden. Diese Annahme wird durch Datensetvisualisierung mittels Dimensionseinbettung gestützt. Dies bedeutet allerdings, dass der Silhouettenkoeffizient fallspezifisch und nur innerhalb des Datensets vergleichbar ist. Auch muss darauf hingewiesen werden, dass die Grenzen zwischen zwei Clustern nicht unanfechtbar sind und dass ein Ort, der einen negativen Silhouettenkoeffizient hat, durchaus besser in ein anderes Cluster passen kann. Generell sollte ein „gutes“ Clustering über wenige Orte mit negativen Silhouetten verfügen.

*Calinski-Harabasz-Kriterium*

Eine weitere Methode, um die intrinsische Qualität eines Clusterings zu überprüfen, ist das Calinski-Harabasz-Kriterium (vgl. Caliński und Harabasz 1974). Dieses basiert wie der K-Means Algorithmus auf der Summe der quadratischen Fehler (SSE). Dabei wird einmal der SSE-Wert für je ein Cluster ($W$)[134] und einmal für den globalen Wert ($B$)[135] berechnet. Der Quotient aus $B$ und $W$ liefert nun einen Calinski-Harabasz-Wert für ein Cluster. Diese Berechnung wird für jedes Cluster wiederholt und aufaddiert. Ein hoher Wert für dieses Kriterium spricht für ein stabiles Cluster. Diese Werte sind von dem Datenset abhängig und somit nicht zwischen verschiedenen Datensets direkt vergleichbar. Ein Vergleich kann allerdings zwischen verschiedenen Clusterparametern $k$ stattfinden. Damit bietet sich die Methode als Entscheidungshilfe für die Wahl der Anzahl der Cluster an. Dieses Kriterium dient zusammen mit dem Silhouettenkoeffizienten als Hauptargument für die Wahl der Cluster bei der Clusteranalyse in Kapitel 4.

*Clusterstabilität*

Die Stabilität eines Clusterings kann überprüft werden, indem man zufällige Teilmengen des Datensets entfernt und anschließend auf Änderung beim Clustering testet. Dabei wird zuerst das eigentliche, vollständige Clustering mit dem Datenset und den zugehörigen Klassen als eine Art *Ground Truth*

134 Engl.: **W**ithin variance.
135 Engl.: **B**asis variance.

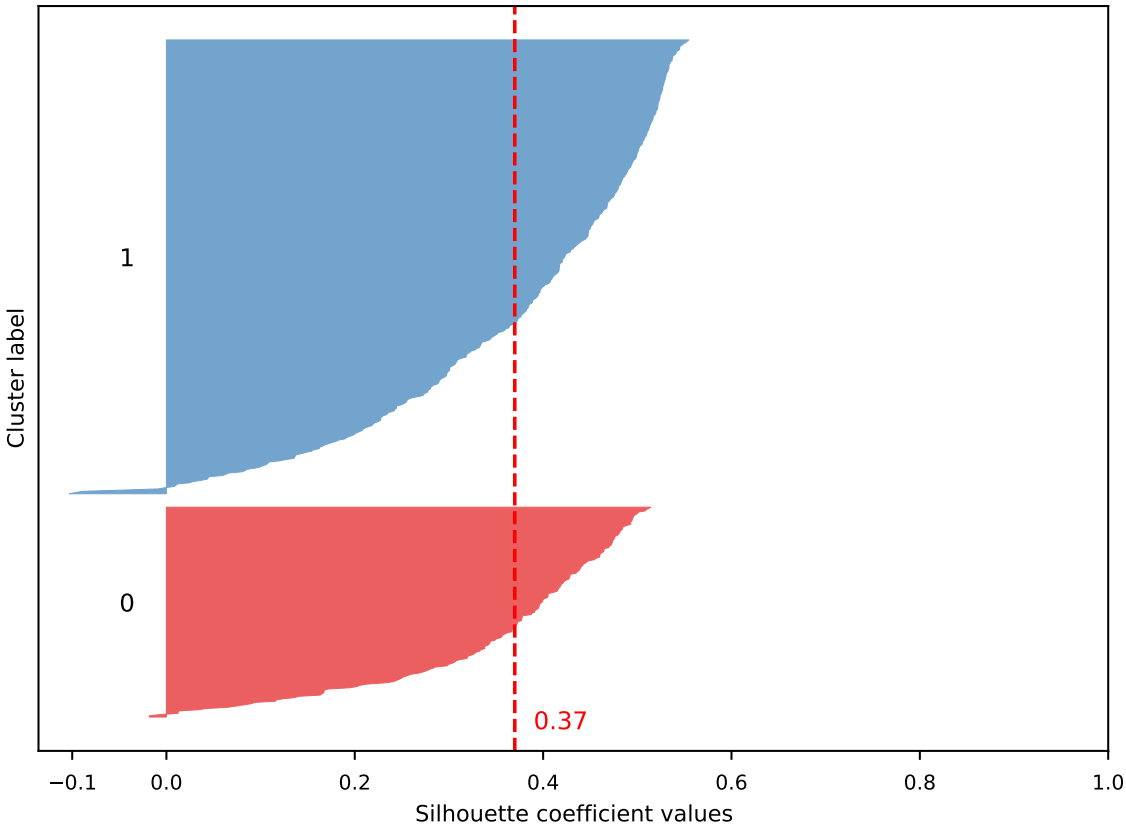



Abbildung 3.5: Beispiel für Silhouettenkoeffizienten basierend auf einem Zweiercluster für die Laute zu mittelhochdeutschen Langvokalen.

beiseite gelegt. Aus diesem Datenset werden nun zufällig ein Trainings- und ein Testset erstellt. Die Trainingsmenge wird mit denselben Parametern wie das originale Datenset neu geclustert. Auf Basis dieser Clusterings werden Klassifikationsalgorithmen trainiert und validiert. Anschließend werden diese Klassifikatoren auf das Testdatenset angewendet. Die Ergebnisse können nun mit den originalen Klassen zu dem Testdatenset verglichen werden. Zusätzlich dazu kann man auch die Ergebnisse des Clusterings anhand des Trainingsdatensets mit den Labels des originalen Datensets vergleichen. Dieser Prozess wird nun mehrmals wiederholt und die Ergebnisse werden zusammengefasst. Für die Bewertung der Ergebnisse gibt es verschiedene Metriken, die aber alle auf der Übereinstimmung zwischen den Labels aus der *Ground Truth* und den Ergebnissen der Klassifikation beziehungsweise des Clusterings arbeiten. Beim Klassifizieren spricht man in diesem Zusammenhang auch von *Genauigkeit*[136], da dies aber für die Bewertung von Clustering ohne externe *Ground Truth* nicht angemessen ist, wird für Clustering der Term *Stabilität* verwendet. Das wiederholte Anwenden von Qualitätskriterien auf Teilmengen des Datensets fällt unter den Sammelbegriff der Crossvalidation. Eine besondere Art der Crossvalidation (*Bootstrapping*) wird im

136 Engl.: Accuracy.

weiteren Verlauf noch genauer vorgestellt. Da Crossvalidieren auf vielen Wiederholungen beruht, ist es sinnvoll, nicht nur die gemittelte Genauigkeit über alle Wiederholungen zu betrachten, sondern sich auch die Streuung der einzelnen Validierungen. Hierbei können wiederum Boxplots hilfreich sein.

Ein Problem beim Clustering ist, dass es bei vielen Algorithmen keine Garantie gibt, dass die Bezeichnung der Cluster bei wiederholtem Anwenden des Algorithmus gleich bleibt. Deswegen kann man beim Überprüfen der Stabilität nicht einfach die Zuordnung der Labels vergleichen, sondern muss vorher überprüfen, ob diese Zuordnung auch kompatibel ist, oder andere Metriken heranziehen. Sollten sich die Cluster gut trennen und relativ stabil gegenüber Änderungen sein, ist es möglich, die Labels mit der gewählten *Ground Truth* zu synchronisieren. Es gibt aber keine Garantie, dass dies immer funktionieren wird. Andere Metriken vergleichen die Cluster durch Mengenübereinstimmung oder durch das Permutieren aller Möglichkeiten. Dabei wird das Problem umgangen, dass die Bezeichnungen der Labels bei vielen Algorithmen zufällig gewählt werden.

Der Adjusted-Rand-Index (vgl. Hubert und Arabie 1985; Rand 1971) ist eine der am häufigsten verwendeten Metriken dieser Art. Dabei wird zunächst der Rand-Index (*RI*) berechnet.

$$RI = \frac{a+b}{C_n^2}$$

Dabei gilt, dass $C$ die *Ground Truth* ist, $a$ alle Elemente bezeichnet, die sowohl in $C$ als auch im Clustering in derselben Gruppe sind, $b$ die Elemente, die in unterschiedlichen Gruppen sind und dass $C_n^2$ alle möglichen Gruppierungen der *Ground Truth* sind. Der Adjusted-Rand-Index *ARI* versucht noch mögliche Zufälligkeiten herauszurechnen, indem ein zufälliger Erwartungswert abgezogen wird.

$$RI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

Der *RI* kann Werte zwischen 1 und 0 annehmen, wobei 1 sehr gut ist und 0 ein zufälliges Clustering bedeutet. Der *ARI* umfasst ein Spektrum von -1 bis 1. Dabei bedeutet 1 wiederum eine vollständige Übereinstimmung und alles kleiner gleich 0 weist auf ein zufälliges Clustering hin.

Eine weitere Metrik ist die Adjusted-Mutual-Information (vgl. Vinh, Epps und Bailey 2010), die auf dem informationstheoretischen Maß der Entropie beruht. Die gemeinsame Information zweier Mengenzuordnungen $U$ und $V$ von $N$ Elementen kann ausgedrückt werden durch:

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_i|}{N} \log\left(\frac{N|U_i \cap V_i|}{|U_i||V_i|}\right)$$

Die Adjusted-Mutual-Information (AMI) rechnet wieder zufällige Verteilungen heraus. Wie der Rand-Index reicht der Ergebniswert von 0 bis 1 beziehungsweise von -1 bis 1 für die AMI. Da die Metriken in dieser Arbeit nicht auf einer verifizierten, sondern auf einer generierten *Ground Truth* beruhen, liefern die Metriken auch eher Ansatzpunkte zur Stabilität als zur „Korrektheit“ des Clusterings.

*Bootstrapping*

Bootstrapping (vgl. Efron 1982) ist eine Crossvalidierungstechnik, die besonders in der Biologie bei der Erstellung phylogenetischer Bäume beliebt ist (vgl. Huson, Rupp und Scornavacca 2010). Diese Technik wurde in der Linguistik bereits von Lameli (2013) bei der Erstellung der Dialekteinteilung angewendet. Bootstrapping basiert auf der Idee des wiederholten Ziehens mit Zurücklegen. Aus einem Datenset wird ein Testdatenset erstellt, indem zufällig Datenpunkte ausgewählt und anschließend wieder zurückgelegt werden, bis so viele Elemente ausgewählt wurden, wie das Datenset insgesamt Elemente umfasst. Anschließend wird ein Klassifikations- oder ein Clusteralgorithmus auf das Testdatenset angewendet. Dieser Vorgang wird sehr oft[137] wiederholt. Aus der Gesamtmenge der wiederholten Anwendungen lassen sich Statistiken ableiten. Man kann zum Beispiel bei einem Clustering die häufigste Zuordnung eines Datenpunktes zu einem Cluster als das dominante Clusterlabel setzen oder sich anschauen, zu welchen Anteilen ein Datenpunkt eine Klasse zugeordnet bekommt. Bootstrapping kann somit als Hilfsmittel zur Berechnung robuster Statistiken eingesetzt werden (vgl. Plonsky, Egbert und Laflair 2015). Im Kontext der Clusteranalyse wird Bootstrapping zur Bewertung der Stabilität der Clusterings eingesetzt. Die Annahme bei Bootstrapping ist, dass sich bei häufiger Wiederholung stabile Cluster herausbilden, obwohl Datenpunkte mehrfach gewählt werden können. Sollte ein Clustering also ein „echtes“ Cluster berechnet haben, dann sollte die Summe der Cluster aus dem Bootstrapping mit den berechneten „echten“ Clustern konvergieren. Anders ausgedrückt, ein Datenpunkt sollte deutlich häufiger das Label des „echten“ Clusters haben als andere Zuordnungen. Diese Methode hat Probleme mit sehr ungleich verteilten Clustern, da die kleineren Cluster häufiger „übersehen“ werden. Bei kleinen Clustern ist dadurch eine höhere Varianz der Clusterzuordnungen anzunehmen. Ähnlich wie bei der Clusterstabilität in Abbildung 3.4 gibt es keine Garantie, dass die Label beim wiederholten Clustering in derselben Reihenfolge zugeordnet werden, deswegen sind direkte Vergleiche zwischen zwei Clusterings nach einem Bootstrapping schwierig[138]. Es lassen sich allerdings die in der Clusterstabilität vorgestellten Metriken anwenden, indem die Hälfte der generierten Cluster als *Ground Truth* gesetzt wird. Wiederum gilt, dass diese Ergebnisse eher als ein Clusterstabilitätsfaktor, als ein solcher der objektiven Korrektheit zu verstehen sind.

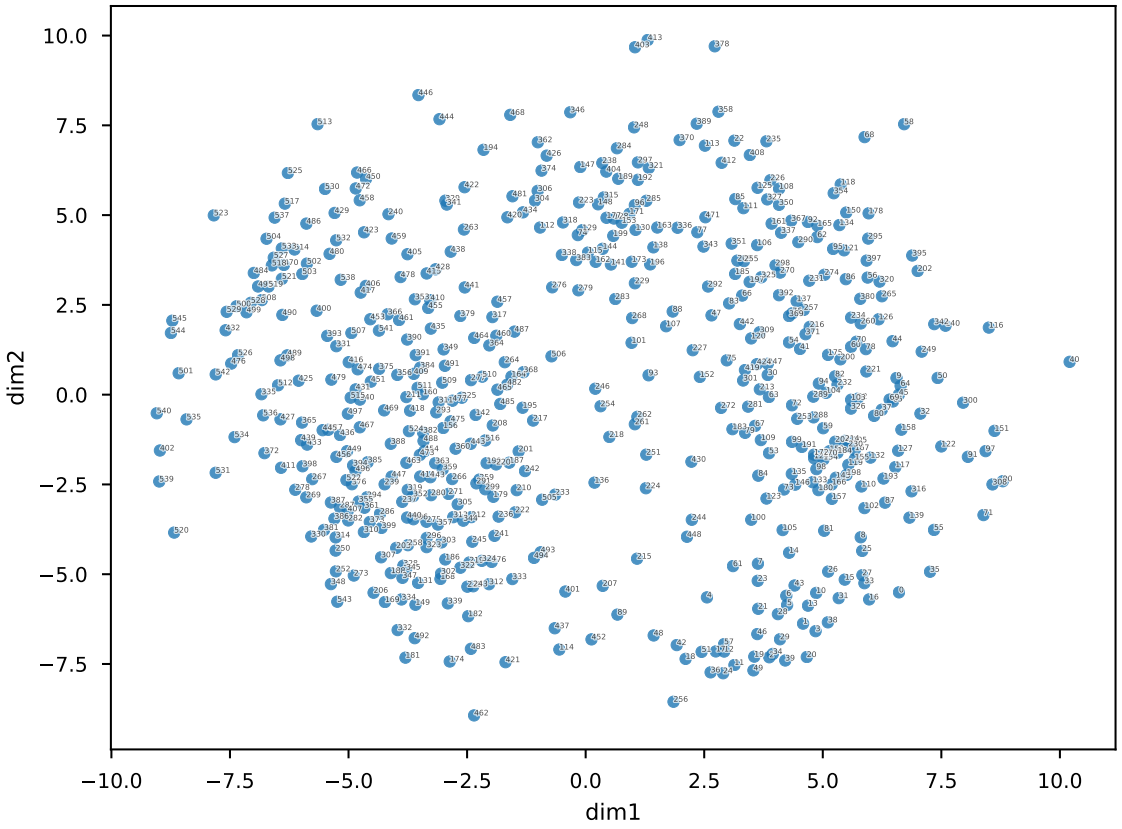
137 Beim Bootstrapping sind 10000 oder mehr Wiederholungen keine Seltenheit. Für diese Arbeit wurden 1000 Iterationen festgelegt.

138 In dieser Arbeit wird dieses Problem abgeschwächt, indem die Cluster von Norden nach Süden renummeriert werden. Der erste Ort und alle Orte, die dasselbe Label wie der erste Ort haben, bekommen immer das Label 0 zugewiesen. Der erste Ort und alle zugehörigen Orte, der nicht zu dem neuen 0-Cluster gehört bekommt das Label 1 und so weiter. Dies bietet keine Garantie, dass die Labelzuweisung immer diskret geschieht, erhöht aber die Chance beträchtlich. Deswegen können auch Karten auf Basis des Bootstrapping erstellt werden.

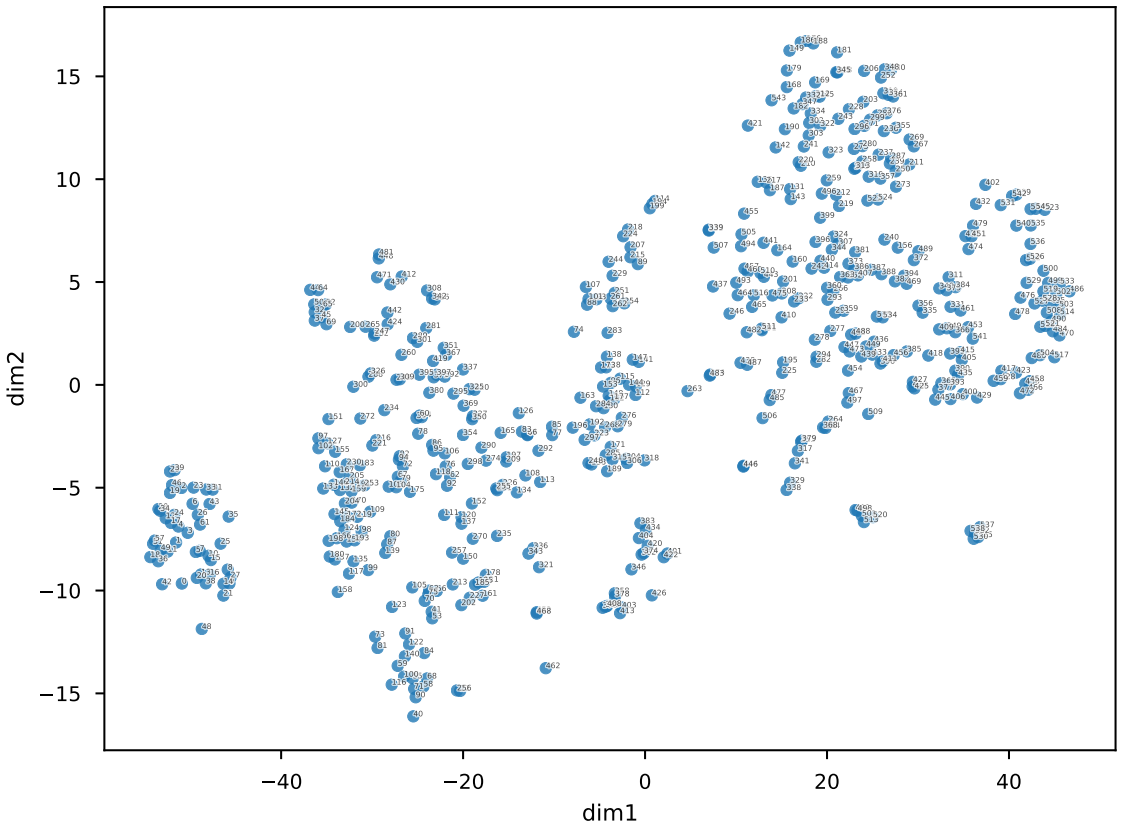
*Dimensionseinbettungen*

Dimensionseinbettungen sollen helfen, eine Vorstellung von der Form der Daten zu bekommen. Da Datensets häufig mehr als drei Dimensionen haben und wir uns höhere Dimensionen nicht vorstellen können, sind Aussagen wie „Das Datenset sollte eine konvexe Form haben" für uns wenig inhaltsreich. Einbettungen sollen eine ungefähre Vorstellung über die Daten vermitteln. Wie bereits erwähnt, ist die Hauptkomponentenanalyse auch eine Art Einbettung und kann bei einer Reduktion auf zwei Dimensionen eine ungefähre Form über die aufgespannte Varianz bieten. Weitere Formen der Einbettung sind die *Multidimensionale Skalierung* (MDS) (vgl. Kruskal 1964) und die *T-Distributed Stochastic Neighbor*-Einbettung (T-SNE) (vgl. Maaten und Hinton 2008). Beide Methoden ermöglichen eine Visualisierung in zwei Dimensionen, allerdings unter sehr unterschiedlichen Gesichtspunkten. Eine MDS versucht die Datenpunkte in einem niederdimensionalen Raum anzuordnen, so dass die relative Entfernung der Punkte untereinander den Entfernungen der Punkte im Ausgangsdatenset entspricht. Da versucht wird, sämtliche Datenpunkte untereinander zu vergleichen, führt dies zu einer globalen Optimierung, wobei jeder Datenpunkt an eine möglichst gute Position im Vergleich zu allen anderen Datenpunkten gesetzt wird. Diese Form der Einbettung ist besonders für Daten mit Raumbezug interessant, weil sie oft mit den räumlichen Positionen korreliert. Die T-SNE ist eine wahrscheinlichkeitsbasierte Einbettung. Sie geht davon aus, dass ähnliche Datenpunkte eine hohe Wahrscheinlichkeit haben, gemeinsam ausgewählt zu werden. Ziel ist es, dass diese Auswahlwahrscheinlichkeit auch bei einer zweidimensionalen Einbettung erhalten bleibt. Dieses Einbettungsverfahren erzeugt stärker abgegrenzte Cluster, allerdings ist die Aussagekraft über die Abstände zwischen diesen Clustern geringer als bei der MDS. Abbildung 3.6 zeigt die Einbettung eines Datensets in die zwei Verfahren. Man erkennt, dass die Einbettung mittels MDS eher einem Raumbild des Untersuchungsgebiets entspricht[139] als die T-SNE, allerdings unterteilt die T-SNE das Datenset in deutlicher unterscheidbare Gruppen.

139 In diesem Fall steht die Einbettung auf dem Kopf. Das kann natürlich passieren, da eine Einbettung keine Aussagen über Norden und Süden treffen kann. Die Rotation der Einbettung ist bei jeder Anwendung des Algorithmus zufällig.




(a) MDS




(b) T-SNE

Abbildung 3.6: Vergleich zwischen einer Einbettung mit MDS (a) und T-SNE (b) angewendet auf das Datenset für phonetische Eigenschaften aller Laute im MRhSA.

# 4 CLUSTERANALYSE AUF DEM DATENSET DES MRHSA

Dieses Kapitel umfasst den Hauptteil der Arbeit, die Clusteranalyse auf dem durch die *phonOntology* generierten Datenset zum MRhSA. Dabei werden die in Kapitel 3 vorgestellten Algorithmen auf das Datenset angewendet. Die Clusteranalyse ist in verschiedene Experimente unterteilt, die jeweils die Daten unter Berücksichtigung eines historischen Lautsystems untersuchen. Das Ergebnis dieser Experimente sind Karten, die die berechneten Cluster in einen räumlichen Kontext setzen und mit den in Abschnitt 2.5 vorgestellten historischen Grenzen vergleichen. Zudem wird analysiert, welche Features für das Zustandekommen der Cluster verantwortlich sind und wie stabil diese Cluster sind. Als „gut" befundene Clusterings können dann als Modell für die Sprachregion oder für einen ausgewählten Aspekt dieser Sprachregion dienen. Alle Clusteranalysen in diesem Kapitel basieren auf den Daten zur älteren Generation (Datenserie 1). Ein Vergleich mit der jüngeren Generation findet in Kapitel 5 statt.

## 4.1 EXPERIMENTE

Die Clusteranalyse zum MRhSA ist in Experimente unterteilt. Ein *Experiment* umfasst die Generierung eines Datensets unter bestimmten Auswahlkriterien (historische Lautklassen) aus dem TripleStore, die Anwendung verschiedener Clusteralgorithmen auf dieses Datenset, die Erzeugung passender Metriken, die bei der Bewertung der Clusteranalyse hilfreich sein können und schließlich eine Visualisierung der Ergebnisse und Metriken in Form von Diagrammen und Karten. Dies geschieht mithilfe eines in Python[140] entwickelten Frameworks, das sehr stark auf einer Funktionsbibliothek für das maschinelle Lernen *skLearn*[141] (vgl. Pedregosa u. a. 2011) basiert. Diese Bibliothek baut wiederum auf dem Goldstandard für numerische Berechnungen in Python auf – *NumPy*[142]. Für eine einfache Bedienung gibt es eine Weboberfläche, entwickelt mit dem React-Framework[143], welches in eine Webanwendung eingebettet ist. Über diese Oberfläche kann ein Experiment erzeugt werden. Experimente werden in einer Datenbank persistent gespeichert. Die Visualisierung der Experimente erfolgt zur Laufzeit, da das Abspeichern sämtlicher Grafiken viel Platz in Anspruch nimmt. Die Experimente lassen sich entweder direkt als eine PDF-Datei oder verteilt auf einzelne PDF-Dateien in einem ZIP-Ordner herunterladen. In diesem ZIP-Order befindet sich außerdem eine automatisch generierte Beschreibung des Experiments mit einer Auflistung der SPARQL-Anfrage, durch die das Datenset generiert wurde und die darauf angewendeten Algorithmen. Das vektorisierte Datenset liegt einmal in unveränderter und einmal in skalierter Form

140 <https://www.python.org>, abgerufen 24.01.2018.
141 <http://scikit-learn.org>, abgerufen 24.01.2018.
142 >http://www.numpy.org>, abgerufen 24.01.2018.
143 <https://reactjs.org>, abgerufen 24.01.2018.

als CSV-Datei bei. Die Ergebnisse des Clusterings sind mit Ortsinformationen versehen, die einen einfachen Import in Geoinformationssysteme, wie zum Beispiel das REDE SprachGIS erlauben. Natürlich kann man auch auf alle Funktionen direkt über Python-Skripte zugreifen, um zum Beispiel weitere Analysen vorzunehmen, die für eine Webanwendung zu viel Zeit in Anspruch nehmen oder zu komplex für eine einfache Parametrierung via Weboberfläche sind.

Die in der Clusteranalyse betrachteten Experimente sind unterteilt nach den historischen Lautklassifikationen und untersuchen, wenn nicht anders erwähnt, alle durch die *phonOntology* inferierten Lauteigenschaften. Die in dieser Arbeit beschriebenen Experimente unterteilen sich in die folgenden:

ALLE: Alle erfassten Observationen.

LANGVOKALE: Alle erfassten Observationen in Karten, die sich mit Phänomenen zu den historischen mittelhochdeutschen Langvokalen befassen.

KURZVOKALE: Alle erfassten Observationen in Karten, die sich mit Phänomenen zu den historischen mittelhochdeutschen Kurzvokalen befassen.

KONSONANTEN: Alle erfassten Observationen in Karten, die sich mit Phänomenen zu den historischen westgermanischen Konsonanten befassen.

Die verwendeten Algorithmen zum Clustering sind mit eindeutigen Kennungen versehen. Der Aufbau einer Kennung ist:

[Kurzname des Algorithmus][Clusteranzahl k]

So steht GMM3 für das Gaussian Mixture Modell mit drei Clustern, KMEANS2 für ein Zweierclustering mit K-Means und WARD5 für ein agglomeratives Clustering mit der Ward-Linkage und fünf Clustern. Für alle Experimente wurden Clusterings mit den drei Algorithmen und einem $k$ von 2 bis 5 durchgeführt. Das bedeutet, jedes Experiment beinhaltet 12 Clusterings mit dazugehörigen Metriken. Da eine vollständige Auflistung aller Modelle viel Platz in Anspruch nimmt, werden im Folgenden nur ausgewählte Clusteranalysen gezeigt. Das Webinterface erlaubt eine beliebige Kombination von historischen Klassifikationen und ein $k$ zwischen 2 und 20 für die drei Algorithmen sowie eine Einschränkung auf die phonetischen Klassen Vokale, Monophthonge, Diphthonge, Konsonanten oder alle Laute. Dies erlaubt die Generierung sehr vieler Experimente. Diese Arbeit beschränkt sich aber auf die oben erwähnten Hauptexperimente.

Die Clusteranalysen sollen eine räumliche Einteilung des Untersuchungsgebiets liefern, die ausschließlich auf den mittels der *phonOntology* annotierten Daten basiert. Dazu gilt es, ein geeignetes $k$ zu finden, das nicht nur eine räumlich sinnvolle Einteilung liefert, sondern auch über ausreichend Stabilität verfügt, um als ein gültiges Clustering akzeptiert zu werden. Eine räumlich sinnvolle Einteilung ist eine Einteilung, die, basierend auf der Annahme, dass „ähnliche“ Datenpunkte auch „ähnlich“ in der Welt angeordnet sind, räumlich zusammenhängende Clusterings in einer Karte erzeugt.

## 4.2 UNTERSUCHUNG ALLER OBSERVATIONEN ZU ALLEN LAUTEN

*Vorverarbeitung*

Das erste Experiment (ALLE) ist eine Analyse aller 773936 inferierten phonetischen Lauteigenschaften der Observationen der Datenserie 1. Die Transformation nach dem One-Hot-Encoding erzeugt ein Datenset mit 47 Dimensionen an 546 Orten. Eine Verteilung der einzelnen Lauteigenschaften auf das Datenset ist in Abbildung 3.1 auf Seite 68 zu sehen. Die Verteilung der Laute erlaubt bereits einige grobe Aussagen über das Datenset. Wie man erkennen kann, gibt es viele Ausreißer in dem Datenset. Die Lauteigenschaften *LateralApproximant* und *Uvular* haben außer den Ausreißern kein erkennbares Datenspektrum. *Unround* hat viele Ausreißer in die negative Richtung und *Round* in die entgegengesetzte Richtung. Ein genauer Blick auf die *LateralApproximant*-Eigenschaft zeigt einen Ausreißer in negativer Richtung und der Median von *Uvular* ist leicht ins Negative verschoben. Dies lässt vermuten, dass *LateralApproximant* weitestgehend gleichverteilt über alle Orte ist und *Uvular* nur an ausgewählten Orten (weniger als der Hälfte des Untersuchungsgebietes) auftritt. Ein Blick in die Datenbank bestätigt diese Vermutung. Der *LateralApproximant* tritt an allen 546 Orten des Untersuchungsgebietes auf, *Uvular* an 187[144].

Daraus kann man schließen, dass der *LateralApproximant* keine besondere Bedeutung für irgendeinen Sprachraum im Untersuchungsgebiet hat, wohingegen das Auftreten des *Uvular* durchaus Einfluss auf einen Raum haben kann, da die Verteilung auf ein paar Orte mit hoher *Uvular*-Frequenz schließen lässt. Der *Round*–*Unround*-Gegensatz bei den jeweiligen Ausreißern lässt auch ein Gebiet vermuten, das durch diesen Gegensatz geprägt wird[145].

Eine genauere Beziehung zwischen den Lauteigenschaften zeigt die Korrelationsmatrix in Abbildung 4.1. Da sich ein Laut aus mehreren Lauteigenschaften zusammensetzt, ist eine positive Korrelation zwischen den lautdefinierenden Eigenschaften zu erwarten. So sind *LoweredClose* und *NearFront* zwei der Haupteigenschaften zur Erzeugung eines [ɪ]-Lautes und haben, wie zu erwarten, eine hohe Korrelation. Auch ist zu beachten, dass Eigenschaften, die zur Erzeugung vieler Laute beitragen, eine Korrelation nahe 0 mit Eigenschaften haben können, wenn der dadurch erzeugte Laut selbst nur eine geringe Varianz besitzt. So wird das [l]-Phon durch *LateralApproximant*, *Alveolar* und *Voiced* beschrieben, man sieht aber eine leicht negative Korrelation zwischen *LateralApproximant* und *Alveolar*. Aus den Beobachtungen zu Abbildung 3.1 weiß man bereits, dass *LateralApproximant* gleichmäßig verteilt ist und nur eine geringe Varianz hat. Die Varianz von *Alveolar* ist aber deutlich größer, deswegen ist die Korrelation zwischen den beiden Eigenschaften gering. Solche Beziehungen helfen beim Identifizieren von Ei-

144 Der Median ist der mittlere Wert einer sortierten Auflistung der Anzahl der skalierten Ausprägungen. Daraus folgt, wenn mehr als die Hälfte einer Ausprägungen 0 sind, der Median 0 ist.

145 Es ist zu beachten, dass diese beiden Eigenschaften unabhängig voneinander modelliert sind und sich nicht notwendigerweise gegensätzlich verhalten müssen, obwohl es der Name nahelegt.

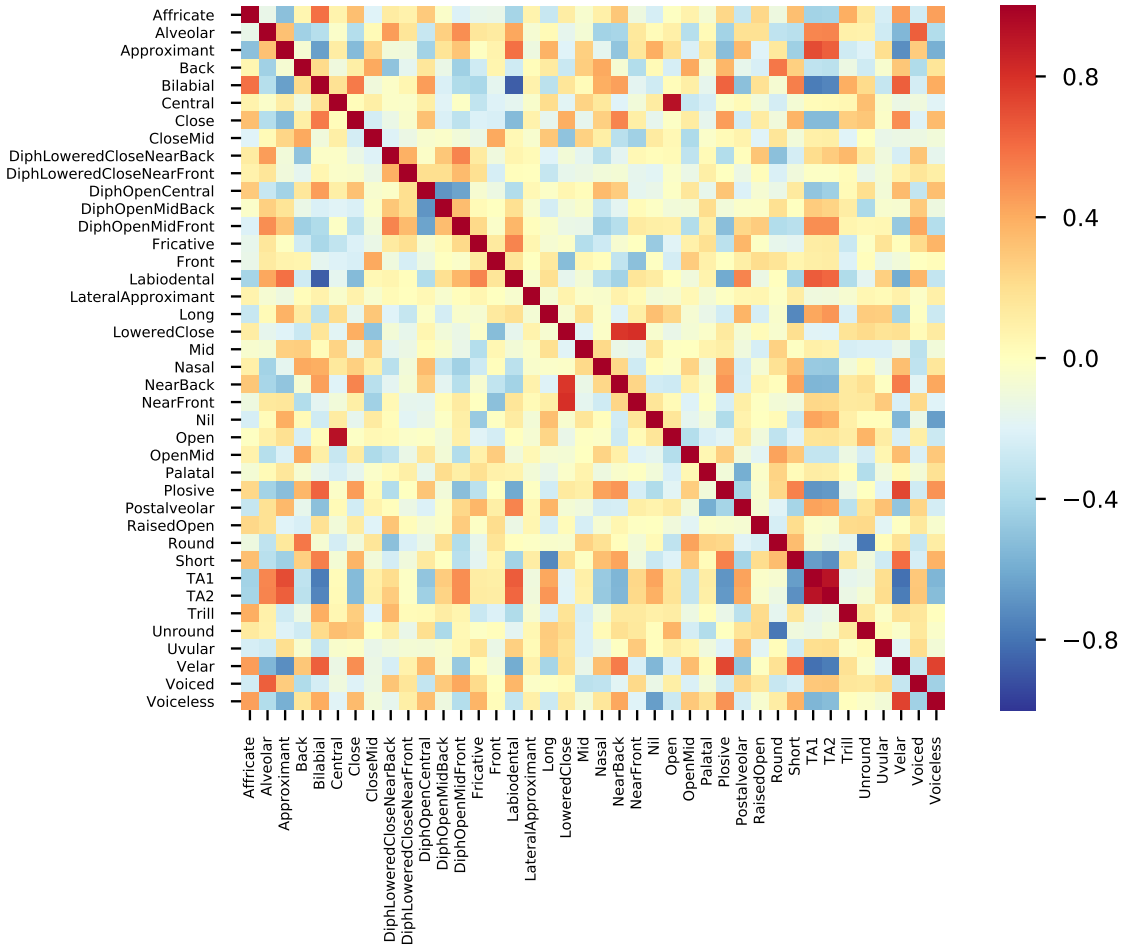



Abbildung 4.1: Die Korrelationsmatrix für alle Lauteigenschaften. Rot eingefärbte Felder bedeuten eine hohe Korrelation, Blau eingefärbte eine negative Korrelation (Antikorrelation). Die Korrelation ist zwischen 1 (vollständige Korrelation) über 0 (keine Korrelation) bis -1 (vollständige Antikorrelation) normiert.

genschaften, die Einfluss auf die Form des Datensets haben. Interessant ist besonders das Zusammenspiel von Eigenschaften, zwischen denen zunächst kein direkter Zusammenhang im Sinne einer Lauterzeugung besteht. Ein solches Zusammenspiel kann auf Sprachraumstrukturen hinweisen. Die Tonakzente, denotiert durch *TA1* und *TA2*, zeigen eine starke Antikorrelation mit der *Short*-Eigenschaft. *Velar* und *Bilabial* korrelieren, zeigen aber deutliche Antikorrelationen mit *Approximant*, *Labiodental*, *Nasal* und den Tonakzenten. Da die Tonakzentgrenze bereits bekannt ist und die Tonakzente nicht direkt aus dem IPA-Vokabular inferiert wurden, sondern als zusätzliche Eigenschaften dem Datenset hinzugefügt wurden, kann man durch eine Korrelation mit den Tonakzenten schließen, in welchem Sprachraum Eigenschaften häufiger oder weniger oft auftreten. Die hoch korrelierten Eigenschaften sollten im Gebiet des Moselfränkischen eine höhere Frequenz haben, die Antikorrelierten im Bereich des Rheinfränkischen. Auffällig

ist, dass die Eigenschaft *Plosive* mit den Tonakzenten antikorreliert, obwohl man durch die *dat/das*-Isoglosse eine umgekehrte Beziehung erwarten könnte. Die *Alveolar*-Eigenschaft korreliert allerdings mit den Tonakzenten. Da diese eine Eigenschaft zur Erzeugung des [t]-Lauts ist, kann angenommen werden, dass zumindest das [t] durch eine hohe Frequenz im MOSELFRÄNKISCHEN zustande kommt.

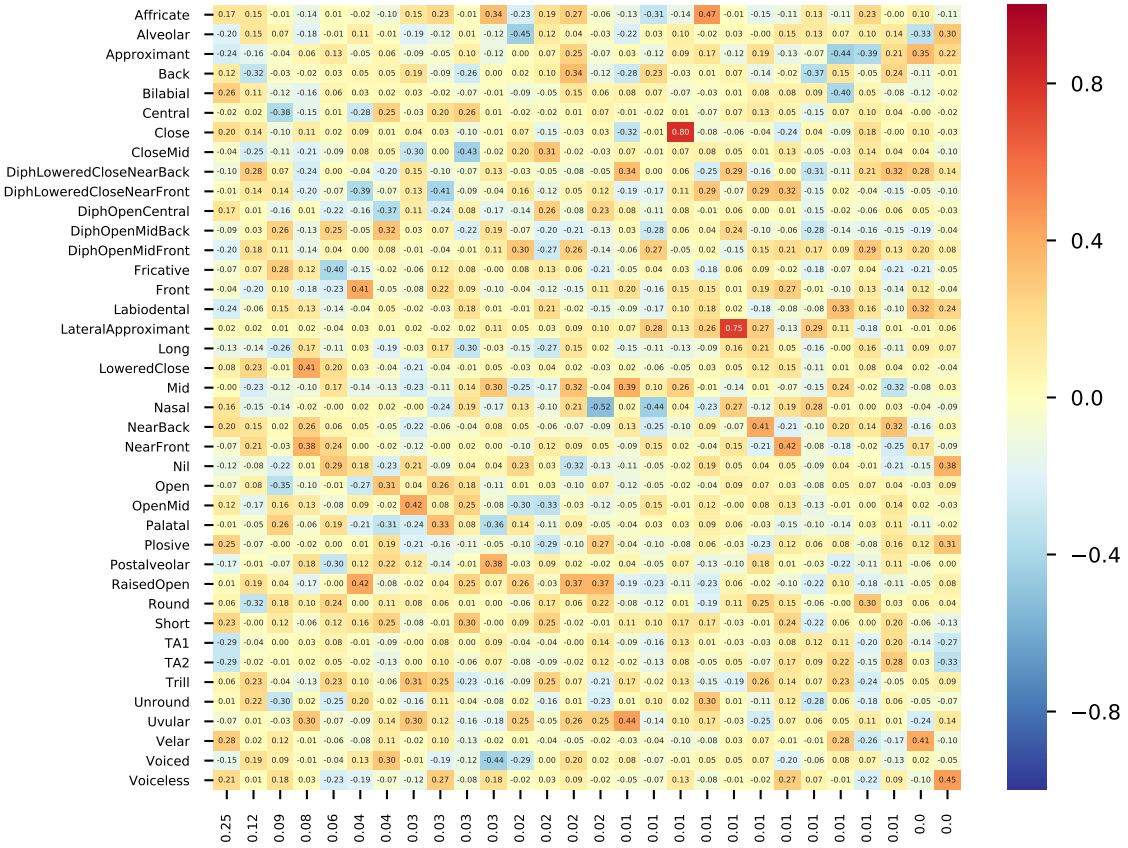



Abbildung 4.2: Anteile der Varianz der ursprünglichen Dimensionen auf die Varianz der neuen, reduzierten Dimensionen nach einer Hauptkomponentenanalyse. Hohe positive (rot) oder negative (blau) Werte zeigen einen hohen Einfluss. Die X-Achse zeigt die erklärte Varianz in Prozent, gerundet auf eine Nachkommastelle.

Die Hauptkomponentenanalyse ist Teil der Vorverarbeitung des Datensets. Dadurch werden die ursprünglich 47 Dimensionen auf 28 reduziert. Diese neuen Dimensionen bilden immer noch 99% der Varianz im Datenset ab. Diese Transformation erlaubt außerdem einen Einblick in die Bedeutung der einzelnen ursprünglichen Dimensionen. So kann man den Anteil der durch die neuen Dimensionen erklärten Varianz der alten Dimensionen in einer Matrixform darstellen. Man sieht in Abbildung 4.2, dass die erste Dimension des neuen Vektors von vielen verschiedenen Eigenschaften beeinflusst wird. Bei den hinteren Dimensionen dominieren häufig wenige Eingenschaften, aber dafür um so stärker. Den Haupteinfluss auf die erste Dimension haben

die Tonakzente, außerdem *Bilabial* und *Velar* für die konsonantischen und *Short*, *Close* und *NearBack* für die vokalischen Eigenschaften.

Mithilfe der Hauptkomponentenanalyse ist es möglich, eine erste Übersichtskarte für das Untersuchungsgebiet zu erstellen. Farberzeugung am Computer erfolgt für gewöhnlich über drei Komponenten, so erzeugt zum Beispiel das **RGB**-Modell Farben aus einer Mischung aus **R**ot, **G**rün und **B**lau Anteilen, das **HSV**-Modell nimmt einen Farbton (**H**ue), eine Sättigung (**S**aturation) und einen Helligkeitswert (**V**alue) zur Farbgenerierung. Man kann also aus den ersten drei Dimensionen des transformierten Datensets einen Farbwert generieren, der einen Datenpunkt repräsentiert. So werden diese Dimensionen auf ein kontinuierliches Farbspektrum abgebildet und es lässt sich über den Farbwert eine Ähnlichkeit zwischen den Datenpunkten (den Orten) visualisieren. Dabei ist zu beachten, dass die ersten drei Dimensionen nur ungefähr 60% der Streuung des ursprünglichen Datenset erklären.

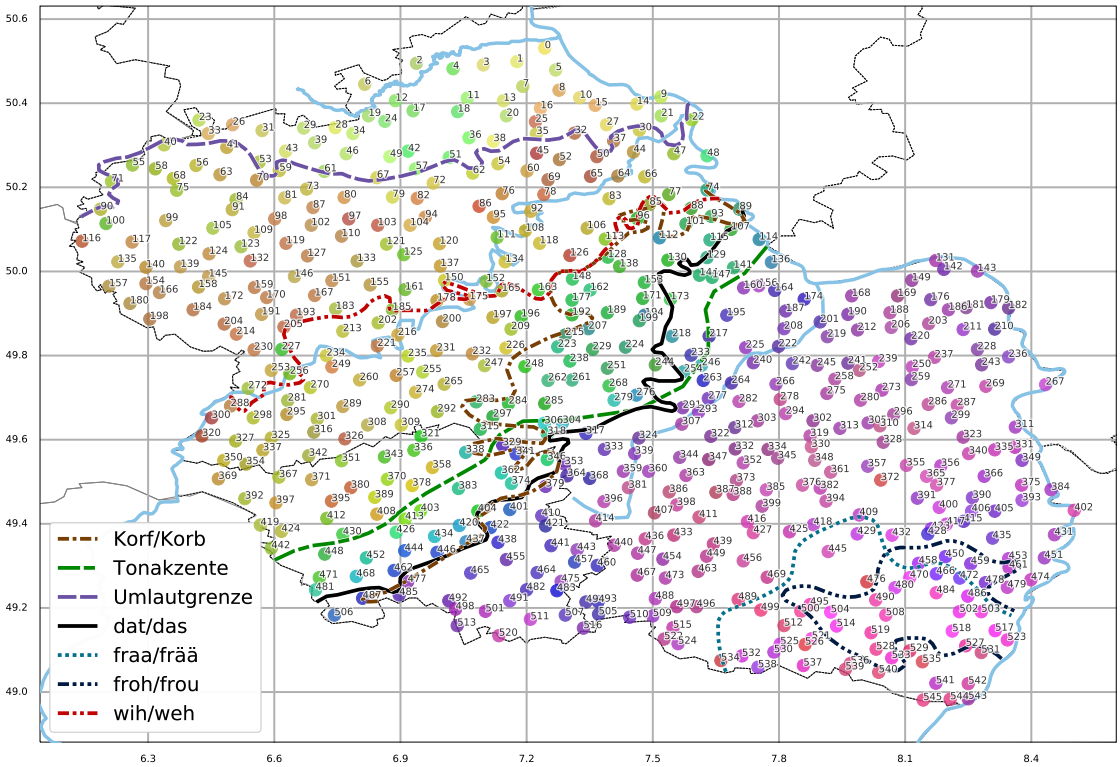



Abbildung 4.3: Übersicht über die Region des MRhSA. Einfärbung der Orte erfolgt basierend auf den ersten drei Dimensionen des PCA-transformierten Datensets für alle Lauteigenschaften und dem HSV-Farbmodell. Dabei sind Sättigung und Helligkeit mindestens 0.5.

Abbildung 4.3 stellt eine derartige Visualisierung dar. Man sieht deutlich eine Trennung zwischen einem lila-rot dominierten Bereich im Rheinfränkischen und einem grün-braun dominierten Bereich im Moselfränkischen. Die Grenze läuft entlang der Tonakzentgrenze und der *dat/das*-Isoglosse. Auffällig ist auch der eher türkisfarbende Bereich zwischen der *Korb/Korf*-Isoglosse und der Tonakzentgrenze. Diese Visualisierung stützt die Behauptung der Zweiteilung des Untersuchungsgebietes in die beiden

Hauptsprachräume entlang der Tonakzent- beziehungsweise der *dat/das*-Grenze und des Vorhandenseins eines Übergangsgebietes entlang dieser Grenzen. Im MOSELFRÄNKISCHEN kann der türkisfarbende Bereich als das Übergangsgebiet interpretiert werden, im RHEINFRÄNKISCHEN der blaue Bereich, der zudem noch eine Häufung im Saarland hat.

*Clusteranalyse*

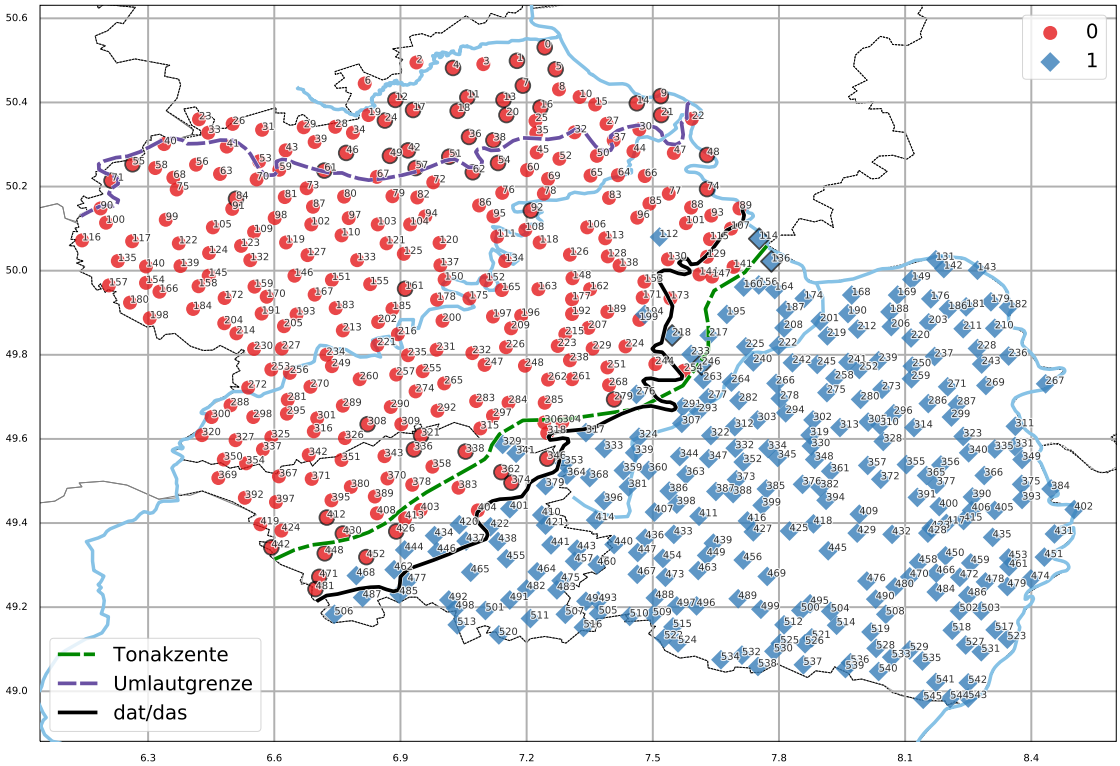
**Box 4.2.1 Bezeichnung der Cluster**

Die Bezeichnung der einzelnen Cluster erfolgt einfach über eine Zahl n ($n \in \{0...k-1\}$ mit $k =$ Anzahl der Cluster). Die spezifischen Cluster werden als n-Cluster bezeichnet; also 1-Cluster für das Cluster mit dem Label 1. Für gewöhnlich erfolgt die Benennung der Cluster von Norden nach Süden. Die Benennung der Cluster ist immer algorithmusspezifisch. Ein 0-Cluster aus einem GMM3 ist nicht dasselbe wie ein 0-Cluster aus einem WARD3. Auch wenn die Cluster häufig mit den in Abschnitt 2.5 vorgestellten Regionen übereinstimmen, ist eine Zuordnung eines Clusters zu einem Sprachraum bereits eine Interpretation.

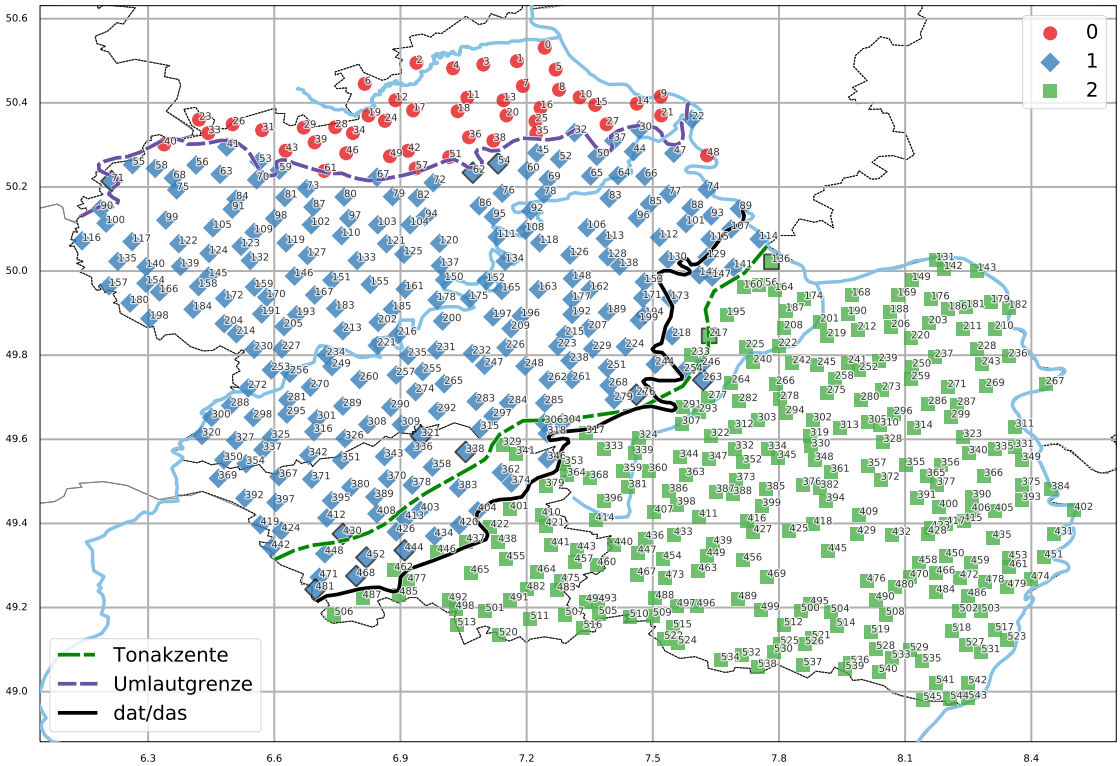
Ausgehend von der Karte in Abbildung 4.3 bietet sich als erste Untersuchung ein Zweiercluster an. Abbildung 4.4a zeigt ein Clustering nach KMEANS2. Man sieht im Bereich nördlich der Entrundungsgrenze und im südlichen Grenzgebiet zwischen der Tonakzent- und der *dat/das*-Grenze eine Häufung von Orten mit negativen Silhouetten. Dies lässt vermuten, dass ein Zweierclustering nicht eine optimale Einteilung der Sprachregion ist und man ein Clustering mit einem größeren *k* versuchen sollte. Die Hauptgrenzen zwischen dem 0-Cluster und dem 1-Cluster sind die Tonakzent- und die *dat/das*-Grenze. Auffällig ist dabei, dass der Grenzverlauf wechselt. Im Norden folgt die Clustergrenze der Tonakzentgrenze, im Süden der *dat/das*-Grenze. Der Wechsel geschieht in der Nähe des Ortes Heimbach (318), das heißt in derselben Region, in der die *Korb/Korf*-Grenze (zu sehen in Abbildung 2.4 auf Seite 62) einen Knick nach Norden macht.

Im Dreierclustering (Abbildung 4.4b) erscheint im Norden ein neues Cluster. Dieses Cluster koinzidiert mit dem UMLAUTGEBIET nördlich der Entrundungsgrenze. Auch stabilisiert sich die südliche Grenze zwischen dem 1- und dem 2-Cluster. Dieses Clustering erzeugt zudem die wenigsten Orte mit negativen Silhouetten. Der durchschnittliche Silhouettenkoeffizient ist mit 0.23 zwar immer noch niedrig relativ zu den Vorschlägen aus der Literatur, dies ist aber für Datensets, die auf einem hoch heterogenen Spektrum basieren, nicht notwendigerweise ungewöhnlich[146]. Insgesamt weisen 14 Orte einen negativen Silhouettenkoeffizienten auf. Der Calinski-Harabasz-Score ist mit 137.28 am zweithöchsten für alle erstellten Clusterings des ALLE-Experiments, was auf ein gutes Clustering für dieses Datenset schließen lässt. Als Stabilitätsmaß wird eine „Genauigkeit“ von 0.91 angegeben

146 Vergleiche dazu <http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html>, abgerufen 20.05.2018.




(a) KMEANS2




(b) GMM3

Abbildung 4.4: KMEANS2 (a) und GMM3 -Clustering (b) für alle phonetischen Eigenschaften. Schwarz umrandete Orte weisen auf eine negative Silhouette hin.

und auch der ARI (Adjusted-Rand-Index)[147] ist mit 0.80 relativ hoch. Das 1-Cluster ist dabei das instabilste. Das GMM3-Clustering kann als eine stabile Einteilung des Untersuchungsgebietes bezüglich aller erfassten phonetischen Eigenschaften gelten.

Das 2-Cluster in Abbildung 4.4b kann damit als Modell für das Rheinfränkische dienen, das 0- und 1-Cluster bilden zusammen das Moselfränkische, wobei das 0-Cluster für sich das Umlautegebiet beschreibt. Auch ein Bootstrapping (siehe Abschnitt A.5, Abbildung A.1a auf Seite 218) zeigt eine deutliche Trennung zwischen den Clustern, die dem Moselfränkischen und dem Rheinfränkischen zugeordnet sind. Allerdings trennen sich das 0- und 1-Cluster nicht so deutlich, wobei eine Präferenz in Richtung der originalen Labelzuordnung weiterhin sichtbar bleibt.

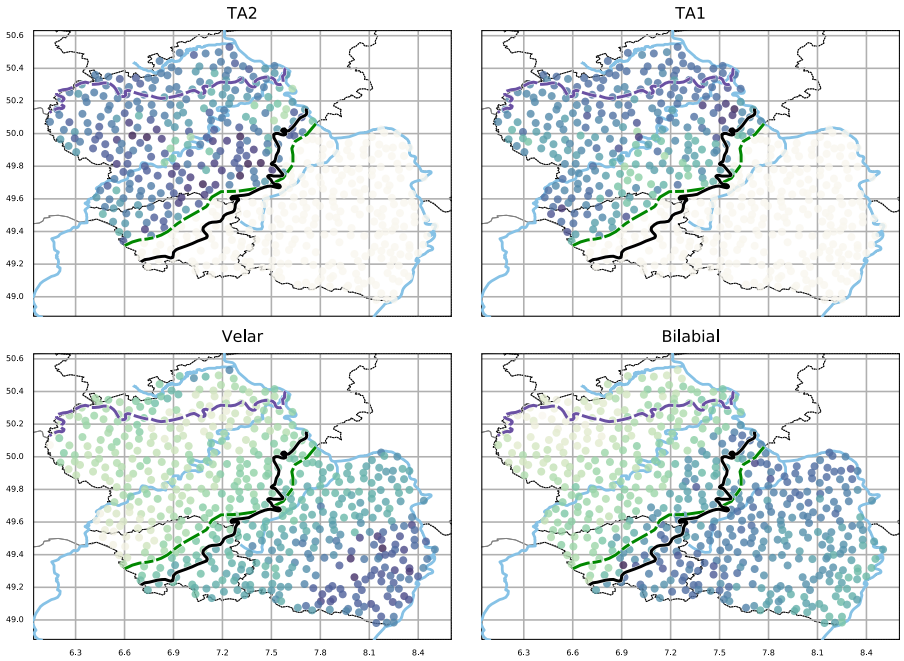
Nachdem durch das Dreicluster eine Einteilung des Untersuchungsgebiets erfolgte, das zumindest nach dem Silhouettenkoeffizienten eine gewisse Stabilität aufweist, gilt es zu untersuchen, welche phonetischen Eigenschaften zur Bildung dieses Clusterings beitragen, wie stabil die Grenzen sind und ob sich noch signifikante Subcluster in den Sprachräumen finden lassen. Mithilfe einer ANOVA[148] (vgl. Weir und Cockerham 1984) lässt sich der Einfluss der einzelnen Dimensionen auf das Modell messen und ein p-Wert für diese Dimensionen bestimmen. ANOVA basiert auf dem *f-test* (vgl. Lowry 2014b) und berechnet den Quotienten des Mittelwerts der Varianz aller Klassen und der Varianz innerhalb einer Klasse für eine Dimension. Als Nullhypothese wird eine zufällige Verteilung angenommen. Tabelle 4.1 zeigt eine Übersicht über die einflussreichsten Dimensionen für ausgewählte Clusterings. Die Tonakzente sind unabhängig vom gewählten Clustering sehr einflussreich. Auffällig ist, dass bei KMEANS2 *Unround* als nicht signifikant eingestuft wird, obwohl es bei höheren Clustern sehr wichtig ist. Das lässt sich dadurch erklären, dass der *Unround–Round*-Gegensatz ein Hauptmerkmal für das Umlautgebiet, also das 0-Cluster in GMM3, ist. Bei einer ANOVA auf dem Zweierclustering wird dieses Merkmal als zufällige Störverteilung im 0-Cluster identifiziert[149] und damit als nicht signifikant. Für eine ANOVA sollten die beobachteten Dimensionen unabhängig und normalverteilt sein, dies ist in diesem Datenset nicht notwendigerweise gegeben. Die Datenbasis ist allerdings ausreichend groß, um trotzdem Aussagen mittels einer ANOVA machen zu können.

Abbildung 4.5 zeigt die Verteilung der vier einflussreichsten Merkmale nach einer ANOVA. Man sieht, dass die Tonakzente eine Art binäres Merkmal sind und in den Clustern, die dem Rheinfränkischen zugeordnet werden, überhaupt nicht auftreten. Eine Übereinstimmung der aus der Datenanalyse erzeugten Grenze und der eingezeichneten Tonakzentgrenze ist zu erwarten, da diese auf denselben Daten basieren. Für das Dreierclustering zeichnet sich zudem der *Round-Unround*-Gegensatz als entscheidendes Merk-

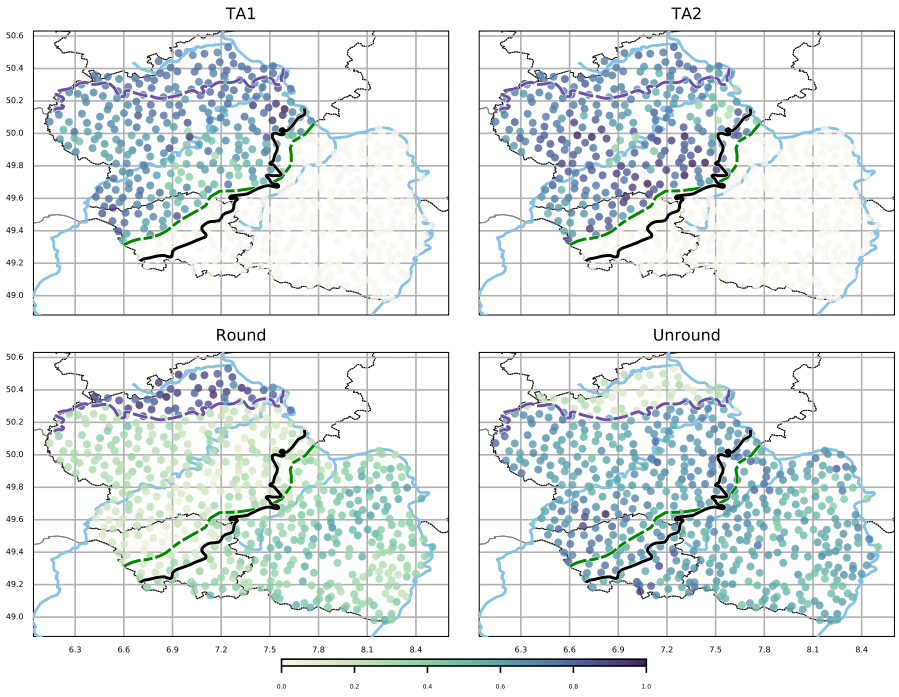
147 Der ARI und andere Stabilitätskriterien basieren auf einer Selbstcrossvalidierung (siehe Abbildung 3.4) und nicht auf einer Ground Truth. Die Werte sind daher mit Bedacht zu interpretieren.

148 **An**alysis **of** **va**riance.

149 Die Algorithmen haben keine Information über die Verteilung der Datenpunkte (Orte) in der Welt und „sehen“ daher nicht, dass diese Punkte ein räumliches Cluster bilden.




(a) KMEANS2




(b) GMM3

Abbildung 4.5: Räumliche Verteilung der vier einflussreichsten Features für KMEANS2 (a) und GMM3 (b) zum dem Datenset für alle phonetischen Eigenschaften. Die Werte sind zwischen 0 und 1 normiert.

Tabelle 4.1: Die zehn höchstsignifikanten ($p$-value < 0.001) und alle nicht signifikanten ($p$-value > 0.05) Eigenschaften für verschiedene Clusterings auf dem Datenset für alle phonetischen Eigenschaften.

| | KMEANS2 | GMM3 | WARD5 |
|---|---|---|---|
| signifikant | TA2 | TA1 | Bilabial |
| | TA1 | TA2 | TA1 |
| | Velar | Round | TA2 |
| | Bilabial | Unround | Labiodental |
| | Plosive | Velar | Velar |
| | Short | Plosive | Round |
| | Labiodental | Bilabial | Unround |
| | Approximant | Short | Approximant |
| | Voiceless | Approximant | Plosive |
| | Alveolar | Labiodental | Voiceless |
| nicht signifikant | Front | Trill | Lateral-Approximant |
| | Unround | Lateral-Approximant | |
| | Trill | DiphLowered-CloseNearFront | |
| | CloseMid | CloseMid | |
| | Palatal | | |
| | Central | | |
| | Mid | | |
| | RaisedOpen | | |
| | DiphLowered-CloseNearFront | | |

mal für das UMLAUTGEBIET im Norden ab (in dem GMM3-Clustering als 0-Cluster bezeichnet).

Um das Zustandekommen der Cluster besser zu verstehen, zeigt Abbildung 4.6 die mittlere Featureausprägung pro Cluster nach dem GMM3-Clustering.

**Box 4.2.2 Interpretationshilfe zu den Grafiken der mittleren Featureausprägung**

Grafiken wie Abbildung 4.6 zeigen die mittlere Featureausprägung über alle gewählten Lautklassen und Cluster. Dabei wird der Mittelwert zu den einzelnen Features über alle Orte eines Clusters gebildet. Der Wer-

tebereich liegt also zwischen dem minimalen und maximalen Wert der skalierten Featureverteilung (zu sehen zum Beispiel in Abbildung 3.1 auf Seite 68). Dabei ist zu beachten, dass ein Wert von 0 der mittleren Verteilung eines Features entspricht. Negative und positive Werte sind also als weniger häufige beziehungsweise häufigere Ausprägungen relativ zu der mittleren Verteilung eines Features zu verstehen. Da die Werte Mittelwerte sind, wird keine Aussage über die Intraclusterstreuung oder über die Häufigkeit eines Features getroffen. Cluster sollten sich paarweise durch negative oder positive Werte für Features unterscheiden.

Man sieht zum einen deutlich den *Round-Unround*-Gegensatz in dem 0-Cluster und die Tonakzente in den dem Moselfränkischen zugeordneten Clustern. Wie bereits in der Korrelationsmatrix angedeutet, ist der *Plosive-Fricative*-Gegensatz relativ zum *dat/das*-Gegensatz vertauscht. In einer Aggregation über alle untersuchten Lautmerkmale ist die Frequenz der *Plosive* im 2-Cluster höher als in den anderen. Für den *Fricative* gilt das Gegenteil, wenn auch bei weitem nicht so deutlich. Generell erscheinen das 1- und das 2-Cluster komplementär zueinander. So hat das 2-Cluster eine hohe Frequenz für *Velar* und *Bilabial*, das 1-Cluster eine niedrige. Für das 0-Cluster zeigen sich neben dem *Round-Unround* Gegensatz als Hauptmerkmal noch weitere Herausstellungsmerkmale gegenüber dem 1-Cluster. So weisen *DiphOpenMidBack*, *Front*, *Back* und *Palantal* eine hohe Frequenz auf, wohingegen *Open*, *Central* und *DiphLoweredNearBack* eine eher niedrige Frequenz aufweisen.

**Box 4.2.3 Interpretationshilfe zu den Grafiken für die Aufteilung nach den historischen Lautklassen**

Diese Grafiken zeigen den Einfluss der einzelnen Lautklassen auf die Features zu einem Cluster. Anders als die Grafiken zur mittleren Aufteilung sind sie nicht um 0 normiert, sondern skalieren zwischen 0 und 1. Die linke Spalte zeigt dieselben Informationen wie die Grafik zur mittleren Ausprägung, nur auf einer 0-1-Skala. Die anderen Spalten zeigen die mittleren Featureausprägungen der entsprechenden historischen Lautklassen, aufgeteilt nach den Clustern. Dabei werden die Daten zu jeder Klasse getrennt auf 0-1 skaliert und *anschließend* der Mittelwert der Cluster berechnet. Da die Verteilungen der Features für jede Lautklasse sehr variieren können, addieren sich die Lautklassen *nicht* zu den mittleren Features (linke Spalte) auf. Eine Featureausprägung einer Lautklasse eines Clusters zeigt zunächst nur an, wie sie im Verhältnis zu den anderen Clustern steht. Allerdings lässt sich aus einem Vergleich zwischen der Ausprägung in einer Lautklasse und der mittleren Gesamtausprägung bewerten, wie sich diese Verteilung zusammensetzt. Zum Beispiel ist in Abbildung 4.6 im 2-Cluster für *wg. t* der *Plosive* mit 0.04 sehr niedrig und der *Fricative* mit 0.72 hoch. Die mittlere Verteilung gibt aber 0.63 und 0.40 respektive an. Im 0- und 1-Cluster sind die Verhältnisse umgekehrt. Dies lässt schließen, dass der saliente *dat/das*-Gegensatz eher wort- beziehungsweise lautklassengebunden ist.

Es ist noch zu beachten, dass Wortklassen, die Featureausprägungen von genau 1 haben, wahrscheinlich über ein sehr homogenes oder sehr geringes Spektrum verfügen. Der Übersicht halber sind Ausprägungen von genau 0 nicht gesondert markiert und einfach weiß belassen. Des Weiteren wird keine Aussage über die Verteilung innerhalb der Orte eines Clusters getroffen.

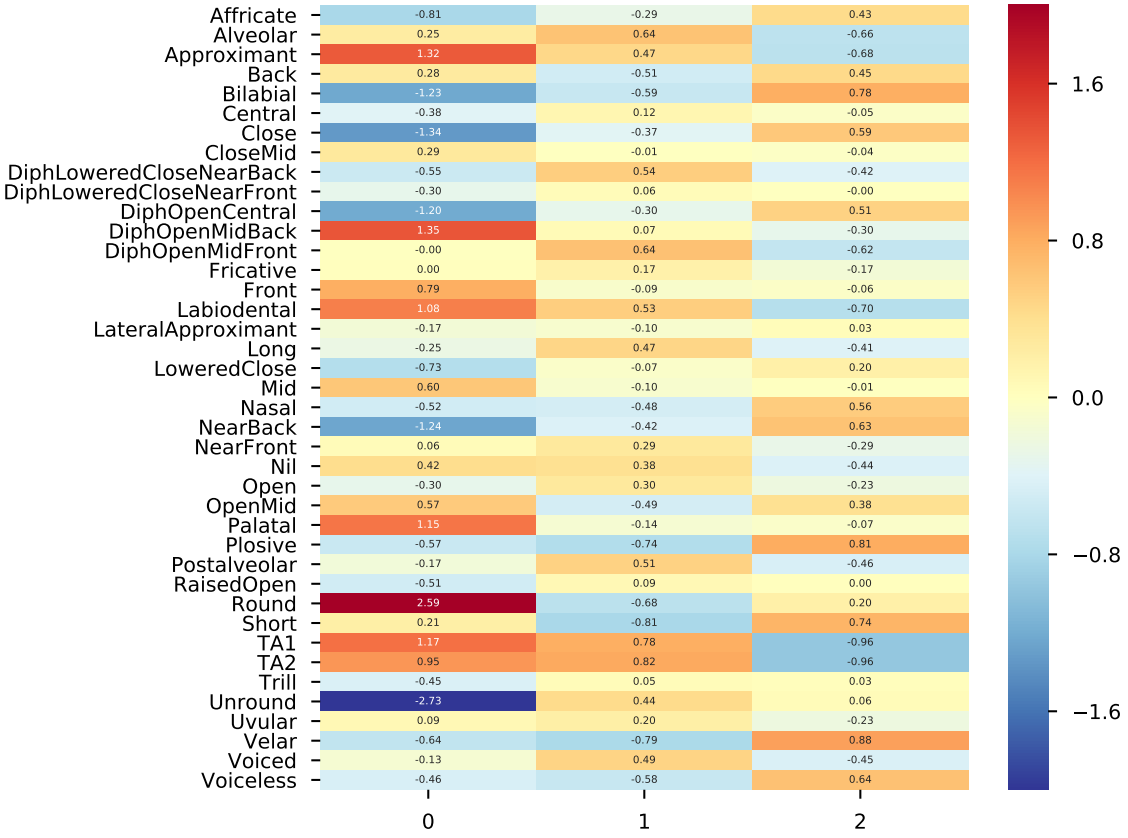



Abbildung 4.6: Mittlere Verteilung der phonetischen Eigenschaften im Datenset zu allen Lauten nach GMM3-Clustering. Rot bedeutet eine hohe Frequenz, Blau eine niedrige.

Eine Aufteilung der Lauteigenschaften nach den historischen Klassen[150] demonstriert den zu erwartenden *Plosive-Fricative*-Gegensatz für das historische *wg. t*, der durch die *dat/das*-Isoglosse markiert wird. Abbildung 4.7 zeigt die Verteilung der einzelnen Klassen auf das historische Westgermanisch. Neben dem *Plosive-Fricative*-Gegensatz in *wg. t* zwischen dem 1-Cluster und dem 2-Cluster sieht man nun, warum dieses Phänomen bei einer Zusammenfassung aller Lautklassen überdeckt wird. In vielen anderen Klassen, zum Beispiel *wg. b*, *wg. g* oder *wg. w*, sind die Auftrittfrequenzen vertauscht, was zur Folge hat, dass die Gesamtfrequenz der *Plosive* im 2-Cluster insgesamt höher ist als in den anderen Clustern. Ob dies ein generelles Phänomen ist oder durch einen Bias innerhalb des Datensets zustande kommt, lässt sich allerdings nicht zweifelsfrei sagen.

150 Aus Gründen der Übersicht werden hier nur die westgermanischen Konsonanten gezeigt.

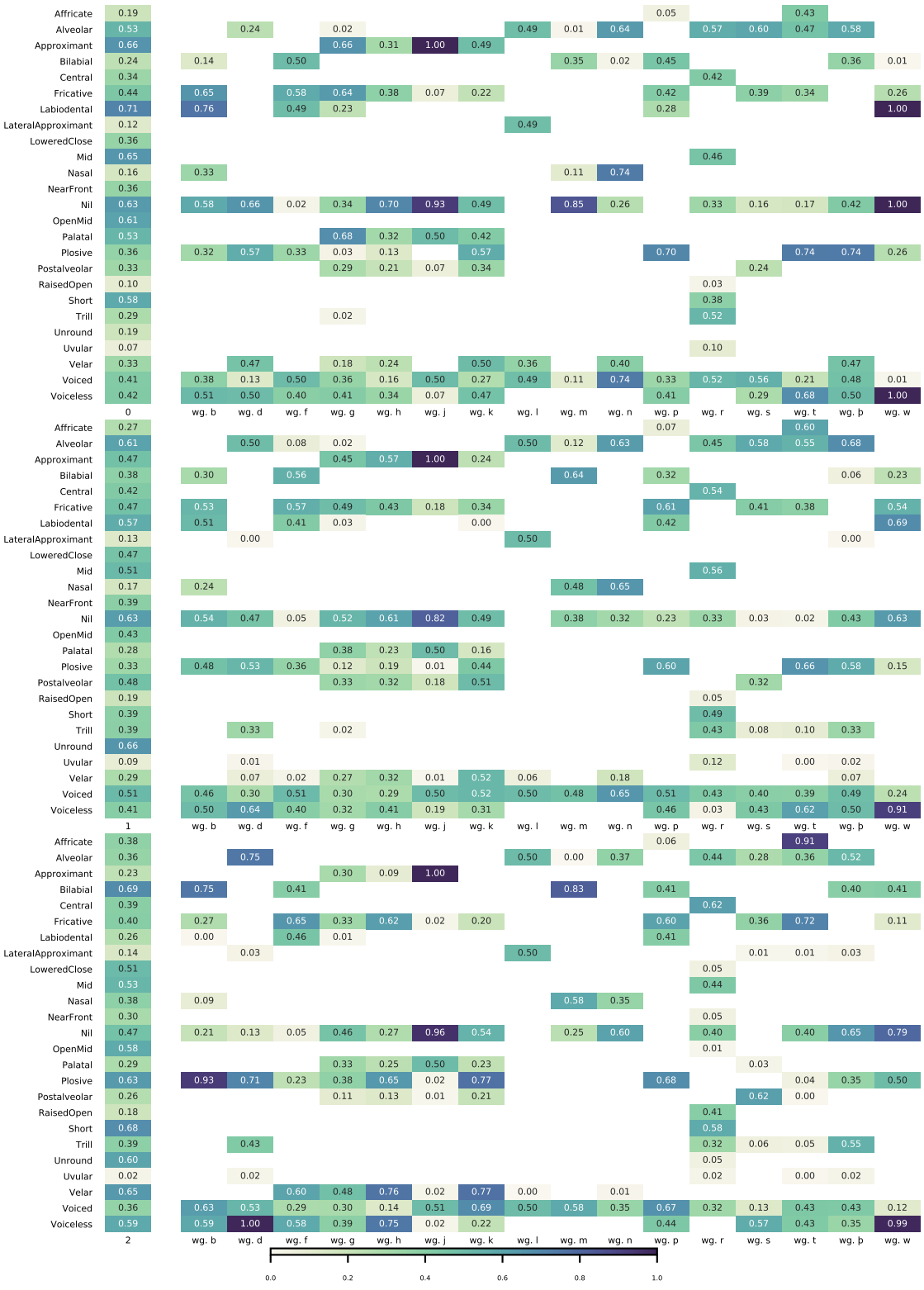



Abbildung 4.7: Ausprägungsverteilung der Cluster nach den einzelnen Lautklassen des Westgermanischen zu GMM3 des ALLE-Experiments.

*Tonakzente*

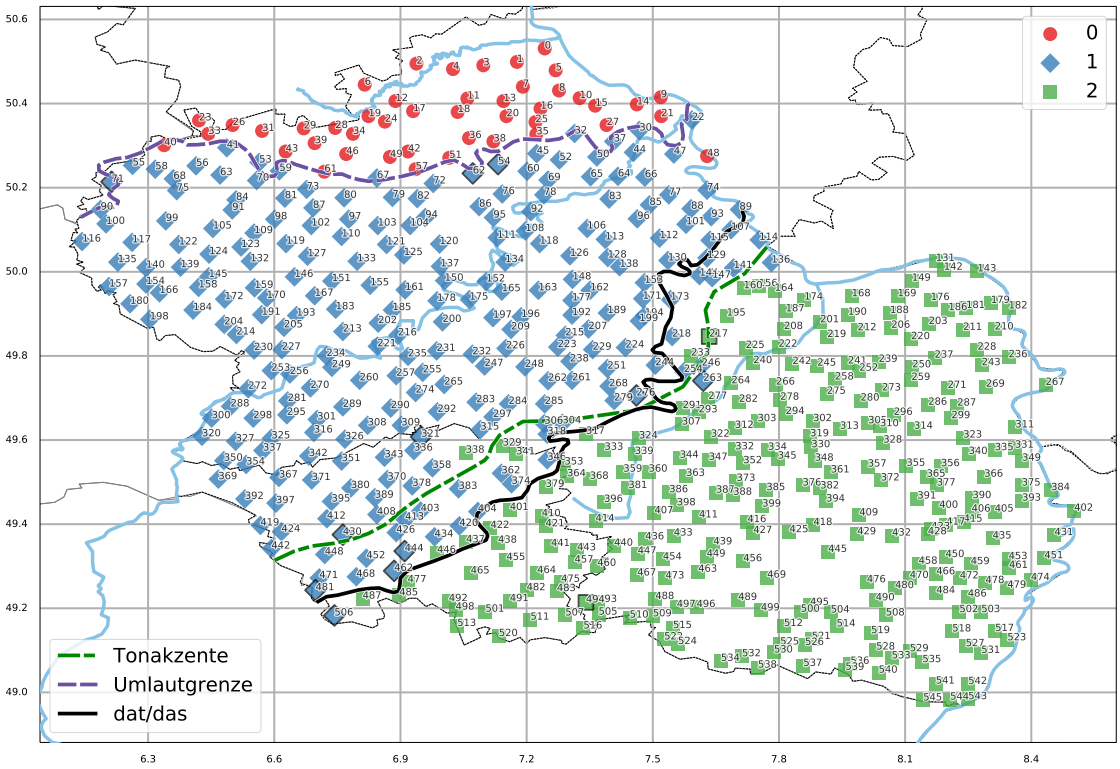



Abbildung 4.8: GMM3-Clustering auf allen phonetischen Eigenschaften ohne Berücksichtigung der Tonakzente.

Die ANOVA bewertet die Tonakzente als sehr einflussreich für das Zustandekommen der Cluster. Da die Tonakzente (TA1 und TA2 zusammen) eine streng binäre Trennung des untersuchten Raumes vornehmen und ausschließlich im Bereich des Moselfränkischen auftreten, ist es zu erwarten, dass sie einen großen Einfluss auf die Form der Cluster haben. Deswegen gilt es zu untersuchen, ob nicht die Tonakzente als Features im Datenset einen zu großen Einfluss auf die Clustererzeugung haben. Dazu wird das Experiment wiederholt, nur dass die Tonakzente herausgefiltert werden.

Ein GMM3-Clustering ohne Berücksichtigung der Tonakzente ist in Abbildung 4.8 zu sehen. Man erkennt, dass die Form der Cluster weitestgehend beibehalten wird. Es gibt kaum Änderungen an den Labelzuordnungen. Insgesamt wechseln nur vier Orte das Label. Lauterbach (506), Püttlingen (462) und Oberheimbach (136) wechseln vom 2-Cluster ins 1-Cluster und Bosen (338) umgekehrt. Der ARI ist mit 0.97 sehr hoch. Der Silhouettenkoeffizient bleibt mit 0.23 unverändert und der Calinski-Harabasz-Wert sinkt mit 133.25 leicht ab. An die Stelle der wichtigsten Merkmale rücken *Round* und *Unround* auf, weiter gefolgt von *Velar* und *Plosive*. Da die Tonakzente nicht unabhängig existieren, sondern auch weitere Lauteigenschaften beeinflussen, sind natürlich Beeinflussungen durch Tonakzenteigenschaften weiterhin in den

Daten vorhanden. Sie sind nur weniger explizit. So antikorrelieren die Tonakzente mit *Short*, und diese Eigenschaft hat im 2-Cluster Gebiet eine hohe Frequenz. Die Tonakzente als eigenständige Features können somit als stabilisierender Faktor angesehen werden, haben als direkte Features aber keinen raumbildverändernden Einfluss auf die Form der Cluster bei GMM3 im ALLE-Experiment.

*Höhere Clusterings*

Im Dreiercluster folgen die Clustergrenzen des 1- und 2-Clusters der *dat/das*-Isoglosse und der Tonakzentgrenze. Das UMLAUTGEBIET im Norden korreliert fast eins zu eins mit dem 0-Cluster. Dies liefert eine gute Einteilung der Region in zwei Hauptregionen, wobei die Hauptregion, die sich mit dem MOSELFRÄNKISCHEN überlappt, noch über eine signifikante Unterregion verfügt. Erhöht man die Clusteranzahl auf fünf, erscheinen weitere Unterregionen. In Abbildung 4.9 werden zwei neue Unteregionen hervorgehoben. Im Gebiet des RHEINFRÄNKISCHEN erscheint am südlichen Rand ein neues Cluster und im Gebiet des MOSELFRÄNKISCHEN bricht ein neues Cluster entlang der Hauptisoglossen heraus. Die Hauptgrenze mit dem 2-Cluster des Dreierclusterings (also das Gebiet des RHEINFRÄNKISCHEN) bleibt bestehen.

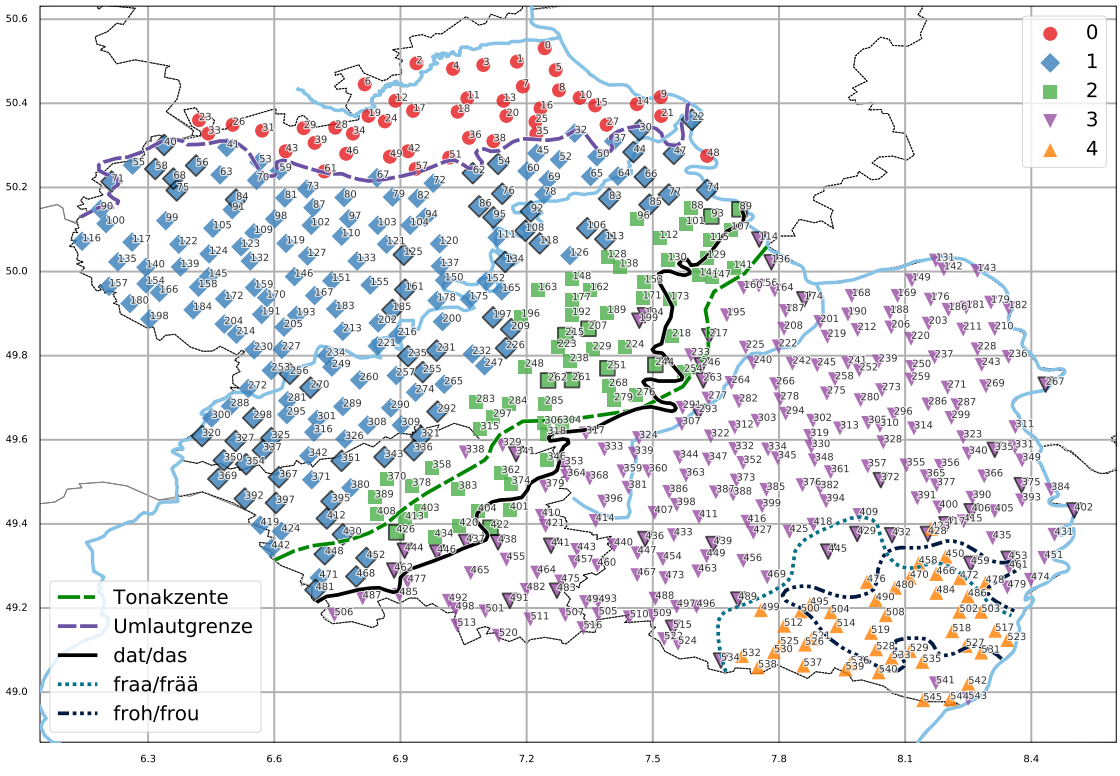



Abbildung 4.9: WARD5 für alle phonetischen Eigenschaften.

Das 4-Cluster markiert den Bereich des SÜDPFÄLZISCHEN RELIKTGEBIETES, wenn auch nicht in völliger Übereinstimmung mit den historischen Struktu-

ren im MRhSA. Das 2-Cluster spannt sich hauptsächlich zwischen der Tonakzentgrenze und der *Korb/Korf*-Grenze auf, reicht allerdings im südlichen Bereich weiter an die Tonaktengrenze heran. Das kleine Gebiet im Süden zwischen der Tonakzentgrenze und der *dat/das*-Grenze teilt sich in 2-Cluster und 1-Cluster. Man sieht jedoch deutlich mehr Orte mit negativen Silhouetten. Insgesamt ist die Stabilität dieses Clusterings deutlich geringer als das Dreierclustering. So beträgt der Silhouettenkoeffizient nur noch 0.12 und auch der Calinski–Harabasz-Wert ist mit 82.49 weit geringer als bei GMM3. Eine genaue Betrachtung der Silhouetten zeigt, dass die Silhouetten für das 0-Cluster und das 4-Cluster hoch sind und der niedrige Gesamtwert besonders durch das 1- und 2-Cluster bestimmt wird. Zwischen diesen beiden Clustern findet keine harte Trennung, wie zum Beispiel zwischen dem 0- und 1-Cluster, statt. Dies spricht für einen eher kontinuierlichen Übergang zwischen diesen beiden Regionen. Ein Bootstrapping des Clusterings zeigt ein ähnliches Ergebnis. Während die 0-, 3- und 4- Labels beim Bootstrapping (siehe Abschnitt A.5, Abbildung A.1b auf Seite 218) weiterhin sehr dominant in ihren Clustern sind, verwischt die Grenze zwischen dem 1- und 2-Cluster.

*Bemerkungen*

In diesem Experiment wurden insgesamt 24 Clusterings erstellt, vorgestellt wurden allerdings nur drei[151]. Die vorgestellten Clusterings zeichnen sich vor allem durch die besten Stabilitätsmetriken für das gewählte k und ein homogenes räumliches Gesamtbild aus. In den meisten Fällen unterscheiden sich die Clusterstrategien nur in einzelnen Orten beziehungsweise in der Anzahl der Orte mit negativem Silhouettenkoeffizient. Es gibt zwei Ausnahmen: zum einen KMEANS3, bei dem sich kein Cluster, das mit dem Umlautgebiet koinzidiert, herausbildet, sondern das 0-Cluster noch Teile der Westeifel abdeckt und das GMM5, bei dem sich ein Cluster um die Tonakzentgrenze herum bildet. Dieses Cluster ähnelt von der Ausdehnung dem Übergangsgebiet in der Wiesingereinteilung (siehe Abbildung 2.3), allerdings besteht es fast ausschließlich aus Orten mit negativen Silhouetten (siehe Abschnitt A.5, Abbildung A.2 auf Seite 219). Es hat sich zudem gezeigt, dass, obwohl die Tonakzente als sehr wichtige Eigenschaft bewertet wurden, sie nicht direkt ausschlaggebend für die Form der Cluster sind und sich zumindest das Dreierclustering auch ohne die Tonakzente rekonstruieren lässt. Als Hauptgrenze wird, basierend auf dem GMM3 Clustering, eine Mischung aus der Tonakzentgrenze im nördlichen Abschnitt und der *dat/das*-Isoglosse im südlichen Abschnitt angenommen.

151 Vier, wenn man den Vergleich mit dem tonakzentlosen Clustering einbezieht.

## 4.3 UNTERSUCHUNG DER OBSERVATIONEN ZU DEN LAUTEN DER HISTORISCHEN KLASSE DER MITTELHOCHDEUTSCHEN LANGVOKALE

*Vorverarbeitung*

Das zweite Experiment (LANG) untersucht die inferierten Eigenschaften zu den Beobachtungen der historischen Langvokale des Mittelhochdeutschen. Insgesamt umfasst dies 147161 Observationen, die wiederum in ein (546, 47)-dimensionales Datenset umgewandelt werden.

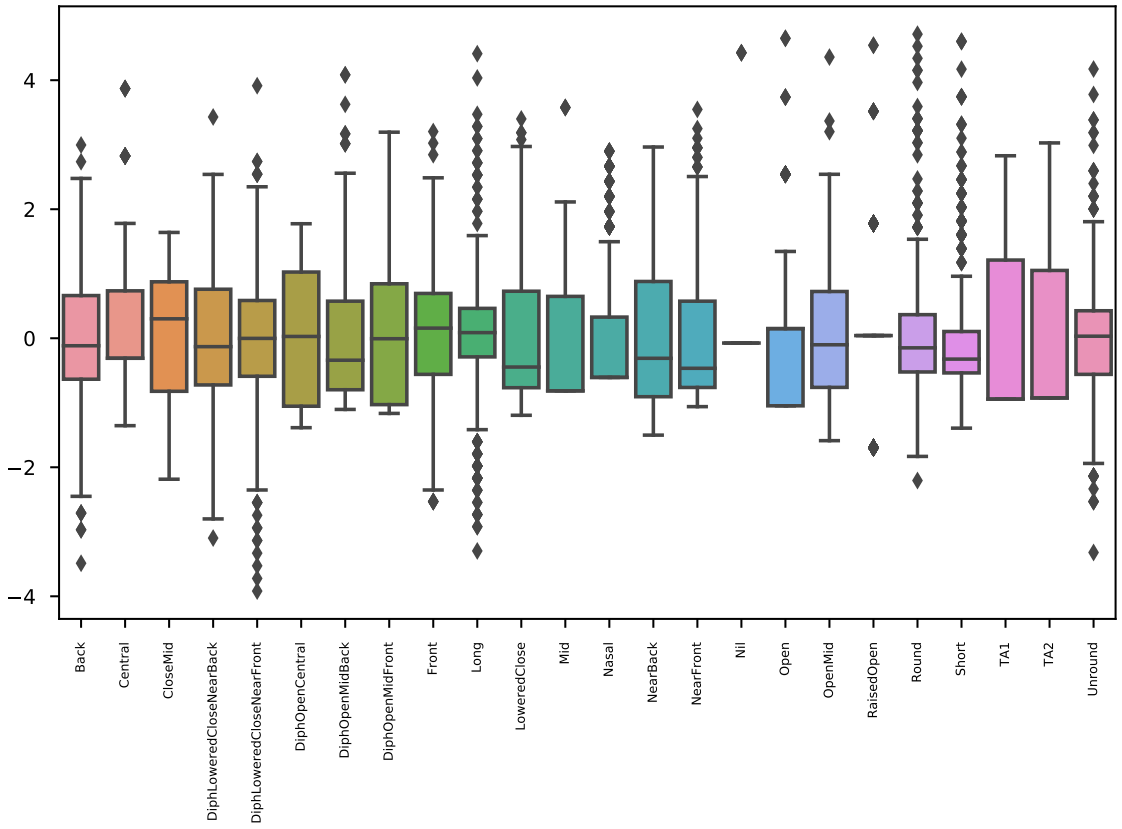



Abbildung 4.10: Verteilung der phonetischen Eigenschaften zu den Observationen der historischen Langvokale des Mittelhochdeutschen.

Bei den historischen Langvokalen treten nicht alle Lauteigenschaften auf, es wird aber zunächst ein vollständiges Datenset konstruiert, damit eine Kompatibilität der Datensets untereinander gewährleistet ist. Dies erlaubt mit gewissen Einschränkungen Matrizenoperationen auf den Datensets der Experimente. Dadurch, dass alle Datensets dieselben Dimensionen und dieselbe Anordnung haben, können nicht überlappende Datensets addiert werden und Teildatensets von einem umfassenderen Datenset, wie zum Beispiel das LANG-Datenset von dem vollständigen ALLE-Datenset, subtrahiert werden. Da das Clustering auf einem dimensionsreduzierten Datenset zum Einsatz kommt, spielen die „leeren“ Dimensionen bei der Weiterverarbeitung

keine Rolle. Auch sind bei den Visualisierungen nur Dimensionen dargestellt, die auch tatsächlich Informationen enthalten.

Da die Langvokale nur eine Teilmenge aller Vokale sind, kann die Verteilung eine andere als die Verteilung aller Lauteigenschaften (siehe Seite 68) sein. Abbildung 4.10 zeigt die Verteilung reduziert auf die historischen Langvokale. Auffällig sind die vielen Ausreißer bei *Long* in beide Richtungen, aber nur in positiver Richtung bei *Short*. Das lässt auf zwei Gebiete schließen. Zum einen ein Gebiet, in dem *Long* ein zusätzliches Merkmal ist und zum anderen ein Gebiet, in dem *Long* durch *Short* ersetzt wird. Der Gegensatz bei den Ausreißern zu *Round* und *Unround* ist bei den Langvokalen nicht mehr so deutlich ausgeprägt wie bei dem Datenset über alle Laute. Tatsächlich haben beide Eigenschaften viele positive Ausreißer. Auffällig ist auch das Vorkommen der konsonantischen Eigenschaft *Nasal*. Auch hier kann man bereits ein räumlich begrenztes Vorkommen inferieren, da der Medianbalken nicht erkennbar ist. Die Eigenschaft kommt damit an weniger als der Hälfte der Orte vor.

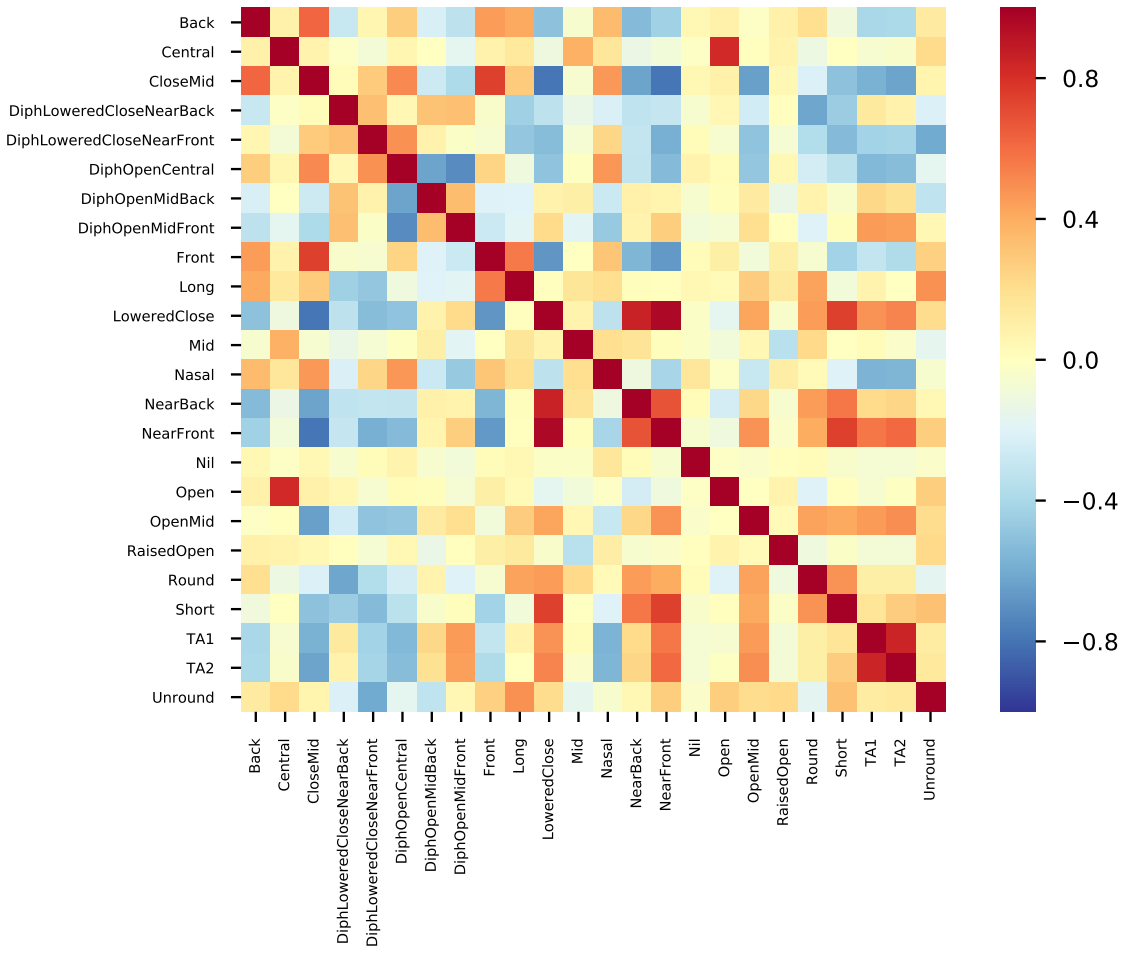



Abbildung 4.11: Korrelationsmatrix zu den Lauteigenschaften der historischen Langvokale des Mittelhochdeutschen.

Die Korrelationsmatrix in Abbildung 4.11 zeigt ein paar Besonderheiten des Langvokaldatensets auf. Auffällig ist die starke Antikorrelation zwischen *Front* und *NearFront* und zwischen *CloseMid* und *LoweredClose*, wenn auch in etwas abgeschwächter Form. Diese beiden Zusammenhänge lassen auf die /e/-/ɪ/-Reihenvertauschung schließen, die Schmidt (2015) beschreibt. Diese Reihenvertauschung ist in der Wenkerkarte „weh“ (WA:113)[152] durch die *wih/weh*-Isoglosse von Trier nach Wiesbaden markiert.

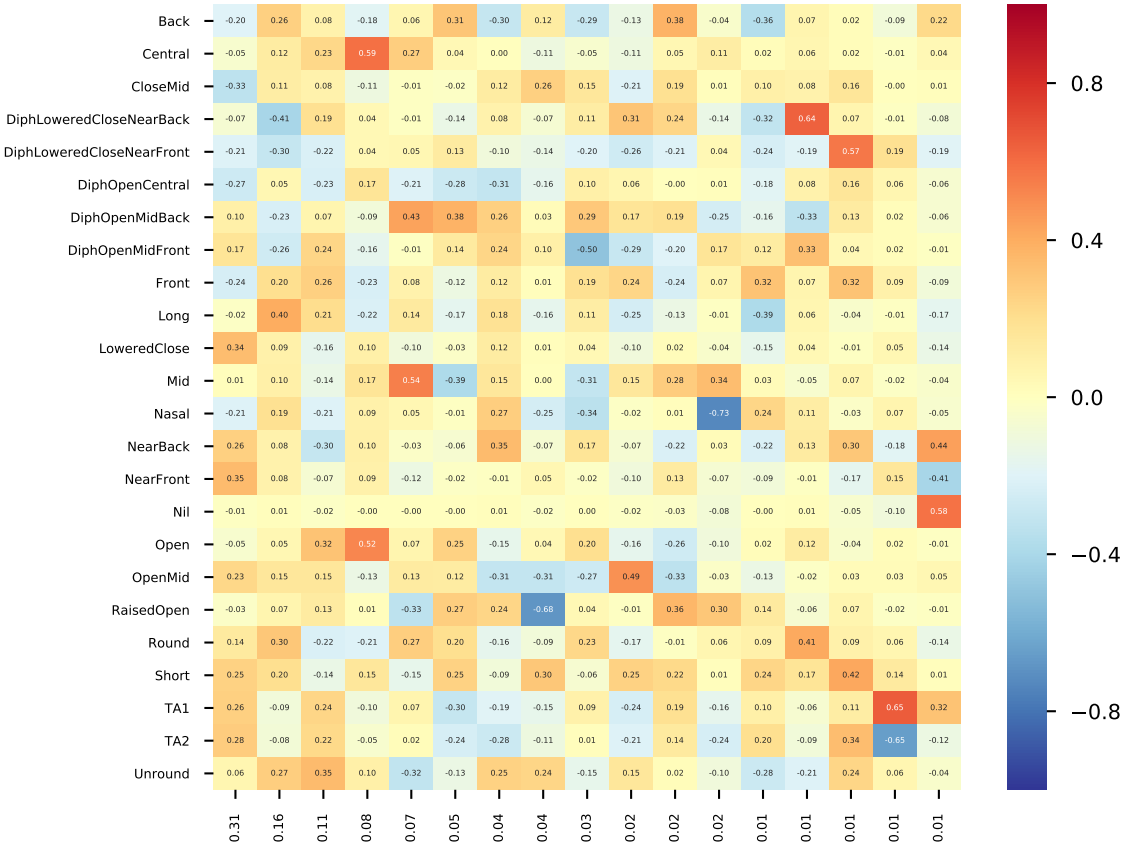



Abbildung 4.12: Anteile der erklärten Varianz der ursprünglichen Dimensionen des Langvokaldatensets auf die Varianz der neuen, reduzierten Dimensionen nach einer Hauptkomponentenanalyse.

Der *Round-Unround*-Gegensatz ist nicht mehr so deutlich wie in der Korrelationsmatrix zu dem ALLE-Datenset (siehe Seite 85). Dies lässt vermuten, dass dieses Phänomen stärker durch die historischen Kurzvokale bestimmt wird. Die hohe Korrelation zwischen *NearFront* und *NearBack* lässt darauf schließen, dass [ɪ]- und [ʊ]-Laute in demselben Gebiet auftreten. Die Eigenschaft *Short* korreliert auch mit *LoweredClose* und *NearFront*, wohingegen *Long* eher mit den Eigenschaften zu [e] korreliert. Dadurch kann das Gebiet, in dem *Short* häufiger auftritt, im [ɪ]-Gebiet verortet werden. Die Antikorrelation der Tonakzente mit *CloseMid* und *Front* lässt vermuten, dass das [e]-Gebiet eher im Rheinfränkischen und das [ɪ]-Gebiet eher im Moselfränkischen zu finden ist.

152 <https://www.regionalsprache.de/SprachGis/RasterMap/wa/113>, abgerufen 29.01.2018.

Eine PCA reduziert das Datenset von den ursprünglich 47 Dimensionen auf 17, wobei die erste und zweite Dimension 31% respektive 16% erklären. Die einflussreichsten Eigenschaften sind wenig überraschend. Dies sind zum einen wieder die Tonakzente, die Eigenschaften des Lautes [e], die negativ gewichtet sind und die Eigenschaften des Lautes [ɪ], die positiv gewichtet sind. Eine komplette Übersicht ist in Abbildung 4.12 gegeben.

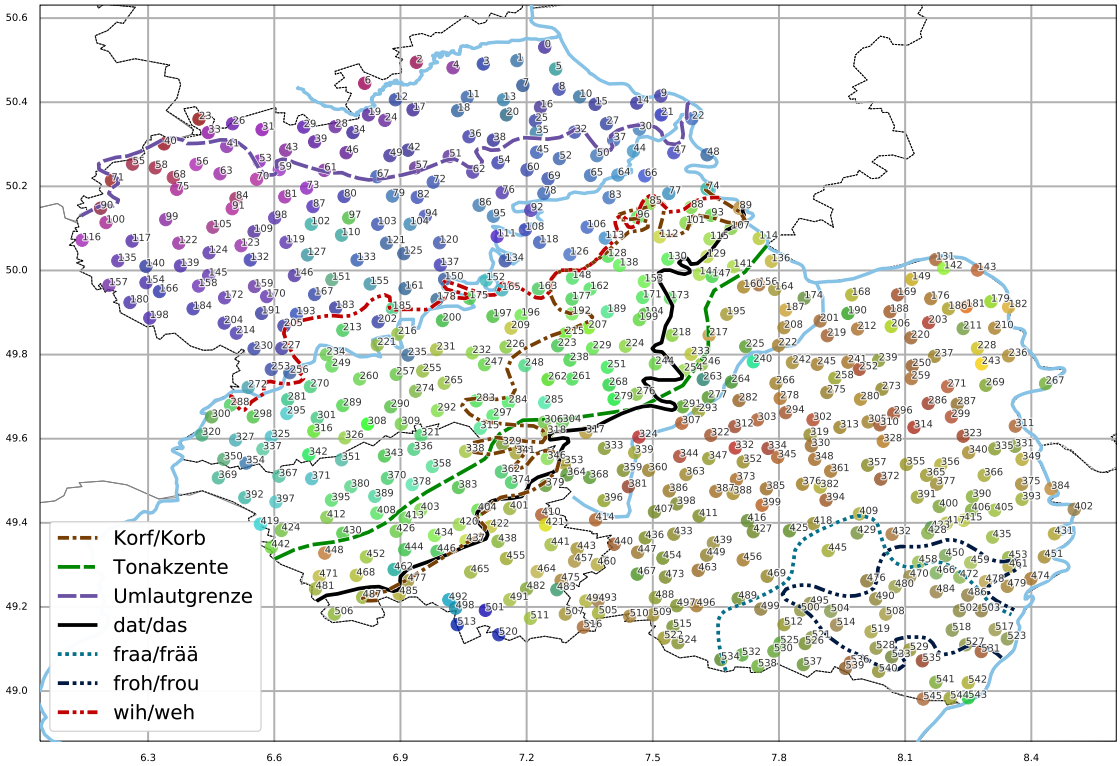



Abbildung 4.13: Räumliche Visualisierung des Langvokaldatensets durch die ersten drei Dimensionen einer PCA, eingefärbt nach dem HSV Farbmodell.

Abbildung 4.13 präsentiert eine Visualisierung der ersten drei Dimensionen des PCA-transformierten Datensets im Raum. Es zeigt sich eine erste Strukturierung der Observationen zu den historischen Langvokalen des Mittelhochdeutschen. Man sieht deutlich den Farbwechsel von blau zu grün entlang der *wih/weh*-Grenze und von dort einen Verlauf ins rötlich-grüne im Bereich des Rheinfränkischen. Auffällig sind die drei dunkelblauen Ausreißer im Süden, namentlich die Orte Eschringen (501), Kleinblittersdorf (513) und Habkirchen (520). Auch gibt es im Nordwesten ein Gebiet, welches eher rot-violett als blau erscheint.

*Clusteranalyse*

Das Zweierclustering nach KMEANS2 ist in Abbildung 4.14 dargestellt. Man sieht deutlich eine Zweiteilung entlang der *wih/weh*-Grenze. Neben vereinzelten Ausreißern, wie die Orte Bleckhausen (97) und Niederkail (151) für

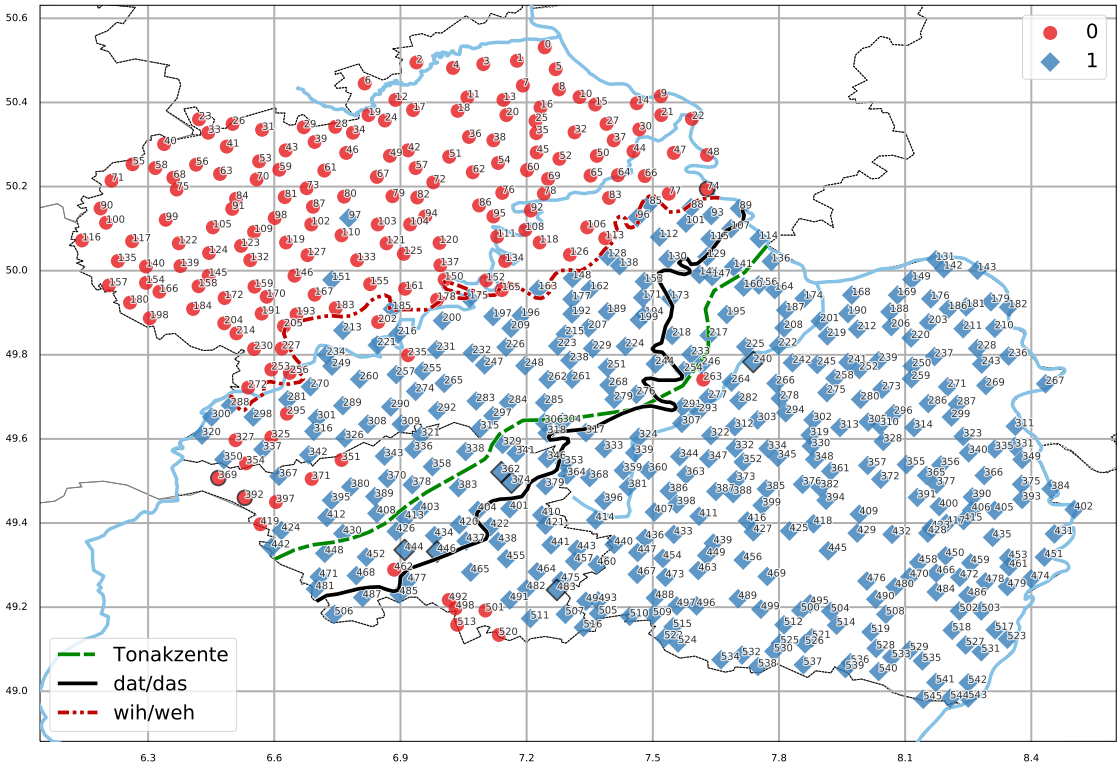



Abbildung 4.14: Clustering der Eigenschaften der historischen Langvokale des Mittelhochdeutschen nach KMEANS2.

das 1-Cluster oder Lauschied (263) und Heidenburg (235) für das 0-Cluster, gibt es zwei auffällige Gebiete. Zum einen ein Gebiet südlich der *wih/weh*-Isoglosse im Westen des Untersuchungsgebietes um den Ort Bethingen (392) und im Süden das bereits erwähnte Ausreißergebiet um Eschringen (501).

Dieses Clustering kann als sehr stabil angesehen werden. Insgesamt gibt es neun Orte mit negativen Silhouetten und der mittlere Silhouettenkoeffizient beträgt 0.37. Der Calinski-Harabasz-Wert ist 337.56, was nicht nur der höchste Wert in diesem Experiment ist, sondern auch deutlich höher als die Werte für andere Clusterings. Es ist zu beachten, dass ein direkter Vergleich, ob ein Clustering „besser“ oder „schlechter“ ist, anhand des Calinski-Harabasz-Wertes zwischen verschiedenen Experimenten nur bedingt aussagekräftig ist. Da aber alle Experimente auf eine ähnliche Datengrundlage zugreifen, kann der Wert als Indikator gesehen werden, wie stabil die Cluster relativ zueinander sind. Der Adjusted-Rand-Index lässt mit einem Wert von 0.98 auch auf ein sehr stabiles Cluster schließen. Ein Bootstrapping (ohne Abbildung) ist so stabil, dass so gut wie alle Orte (98%) zu 100% ausschließlich ihrem Cluster zugewiesen werden[153]. Insgesamt kann daraus geschlossen werden, dass sich diese beiden Cluster deutlich unterscheiden.

153 Es ist zu beachten, dass dies bei einem k von 2 und ausreichend großen Clustern nicht ungewöhnlich sein muss, insbesondere wenn sich diese Cluster klar unterscheiden. Bootstrapping auf höheren Clusterings sollte mehr Varianz aufweisen.

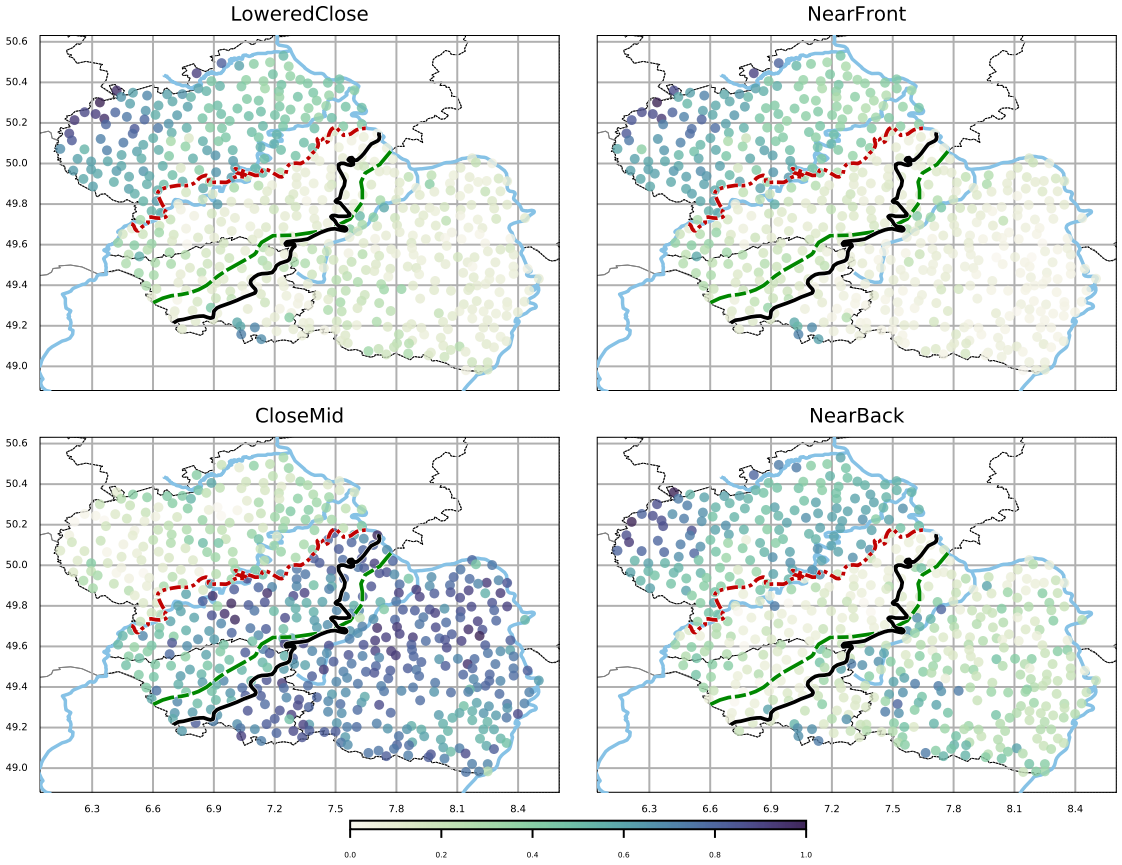



Abbildung 4.15: Räumliche Verteilung der vier einflussreichsten Features für KMEANS2 zu dem Datenset zu den historischen Langvokalen.

Abbildung 4.15 und die Spalte KMEANS2 in Tabelle 4.2 zeigen die Bedeutung der einzelnen phonetischen Eigenschaften für dieses Clustering. In den Karten sieht man, dass die Features die für die Erzeugung von [ɪ]- und [ʊ]-Lauten im Gebiet nördlich der *wih/weh*-Grenze deutlich häufiger sind als im Süden, wohingegen die Features des [e]-Lautes (hier nur *CloseMid* dargestellt) im Süden deutlich frequenter sind. Dies liefert einen Hinweis auf die Vertauschung der Langvokalreihen von /e/ ~ /o/ zu /ɪ/ ~ /ʊ/ im nördlichen Moselfränkischen, da die für die entsprechenden Laute verantwortlichen Eigenschaften als statistisch höchst signifikant für das Zustandekommen der Clusterings bewertet werden, welche sich entlang der historischen *wih/weh*-Isoglosse trennen.

Abbildung 4.16 dokumentiert die Aufteilung der Laute nach den einzelnen Lautklassen. Hier zeigt sich deutlich die Reihenvertauschung in *mhd. ê* und *mhd. ô*. Im o-Cluster dominieren stark die [ɪ] und [ʊ] erzeugenden Eigenschaften, während im 1-Cluster die Eigenschaften von [e] und [o] deutlich häufiger vorkommen. Für das o-Cluster ist auffällig, dass sowohl *NearBack* als auch *NearFront* bei *mhd. ô* sehr dominant sind, was auf eine Teilung der Ausprägung der Lautklasse in [ɪ] und [ʊ] schließen lässt. Die Ausprägungen zu *mhd. œ* sind ähnlich wie die zu *mhd. ê*. *Mhd. æ* ist auch ähnlich zu *mhd.*

*ê*, weist aber zusätzlich noch die Eigenschaften von [e] auf. Im 1-Cluster fallen auch die Eigenschaften zu *mhd ê*, *mhd. æ* und *mhd. œ* als [e] zusammen. *Mhd. î* und *mhd. iu* fallen im 1-Cluster als [a͡ɪ]-Diphthong zusammen, während im 0-Cluster *mhd. î* eher als [ɛ͡ɪ]-Diphthong vorkommt und *mhd. iu* eine deutlich breitere Varianz an Ausprägungen hat. Beide Cluster haben eine merkliche Ausprägung der Eigenschaft *Long*, das 0-Cluster hat zusätzlich noch *Short* als erhöhte Eigenschaftsausprägung, was auf Subregionen mit *Short* als prägendem Merkmal hinweist.

Ein genaues Betrachten dieser Abbildung erklärt auch die Zuordnung des Ausreißergebietes um Eschringen (501) zu dem 0-Cluster. In der historischen Lautklasse *mhd. iu* sind auch die Eigenschaften *LoweredClose* und *NearFront* etwas stärker ausgeprägt. Die Lautklasse *mhd. iu* enthält Wörter, wie „Feuer", „Leute" oder „heute", also Laute, die im Standarddeutschen mit einem [ɛ͡ʊ]-Diphthong realisiert werden. Ein Blick in historische Sprachkarten, wie zum Beispiel die Wenkerkarte zu „Feuer" (WA:77)[154], zeigt das Gebiet um Eschringen (501) als Teil eines Gebietes, in dem die Diphthongierung nicht stattgefunden hat und *mhd. iu* mit einem /ɪ/ realisiert wurde. Sprachhistorisch gehört dieses Gebiet zum Alemannischen. Da das Clustering auf Basis einer Zusammenfassung der historischen Lautklassen geschieht, wird dieses Gebiet mit zu dem 0-Cluster gezählt, da dort dieselben Eigenschaften vorkommen. Dies stützt die Behauptung, dass der /ɪ/-/e/-Gegensatz entscheidend für die Clusterbildung bei einem Zweierclustering ist, da der Algorithmus auf eine hohe Frequenz von [ɪ]-Eigenschaften reagiert.

Das kleine 0-Cluster-Gebiet um den Ort Freudenburg (354) am südwestlichen Ende der *wih/weh*-Grenze wird durch *mhd. ô* und *mhd. œ* geprägt. Für *mhd. ô* folgt es der Lautvertauschung von [o] zu [ʊ], wie auch im Hauptgebiet. Bei *mhd. œ* haben wir eine starke Ausprägung von [ɪ]-Eigenschaften, was wiederum mit [ɪ]-Dominanz des 0-Clusters koinizidiert.

Die einzelne konsonantische Eigenschaft *Nasal* tritt im 0-Cluster überhaupt nicht auf. Ob die Eigenschaft allerdings noch dem Moselfränkischen oder dem Rheinfränkischen zugeordnet ist, lässt sich aus dem Zweierclustering nicht ablesen, da das 1-Cluster deutlich über die Hälfte der Orte umfasst.

*Höhere Clusterings*

Ein Blick auf höhrere Clusterings zeigt weitere Grenzen und Phänomene. Das Dreierclustering in Abbildung 4.17a schildert wieder eine Aufteilung des Untersuchungsgebietes in einen moselfränkischen Teil mit dem 0- und 1-Cluster und ein rheinfränkisches Gebiet mit dem 2-Cluster. Interessanterweise verläuft die Grenze zwischen dem 1-Cluster und dem 2-Cluster nicht vollständig an der Tonakzentgrenze, sondern ähnlich wie die Grenze im Clustering über die Gesamtdaten (siehe Seite 89), im nördlichen Grenzbereich entlang der Tonakzentgrenze, im südlichen entlang der *dat/das*-Grenze. Auffällig sind die beiden Häufungen von Orten mit negativen Silhouettenkoeffizenten im 2-Cluster: einmal bei Orten in der Nähe des Ortes Niederwürzbach (482) nahe der *dat/das*-Isoglosse und bei Orten in der Nähe des Ortes

154 <https://www.regionalsprache.de/SprachGis/RasterMap/wa/77>, abgerufen 30.01.2018.

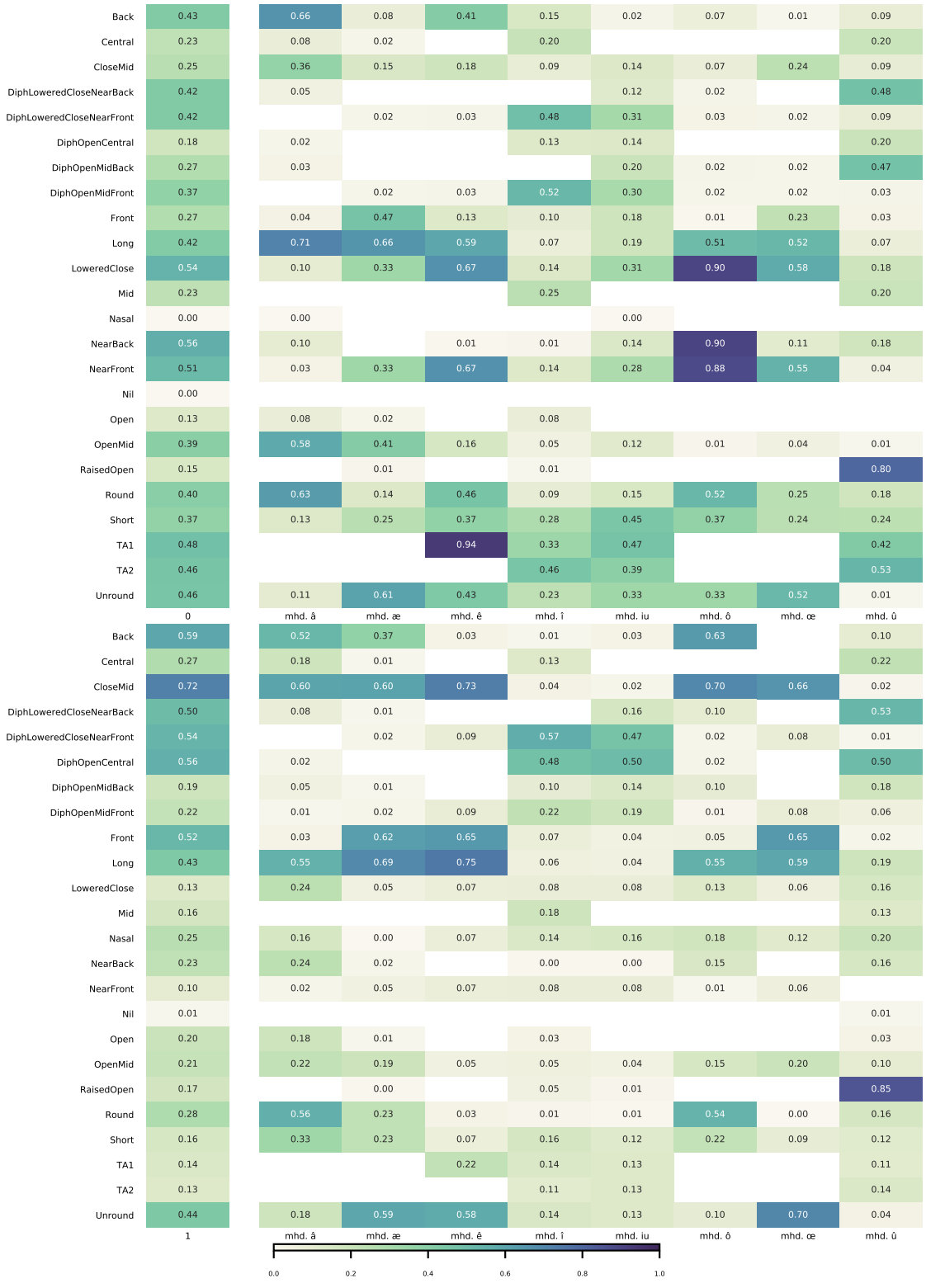


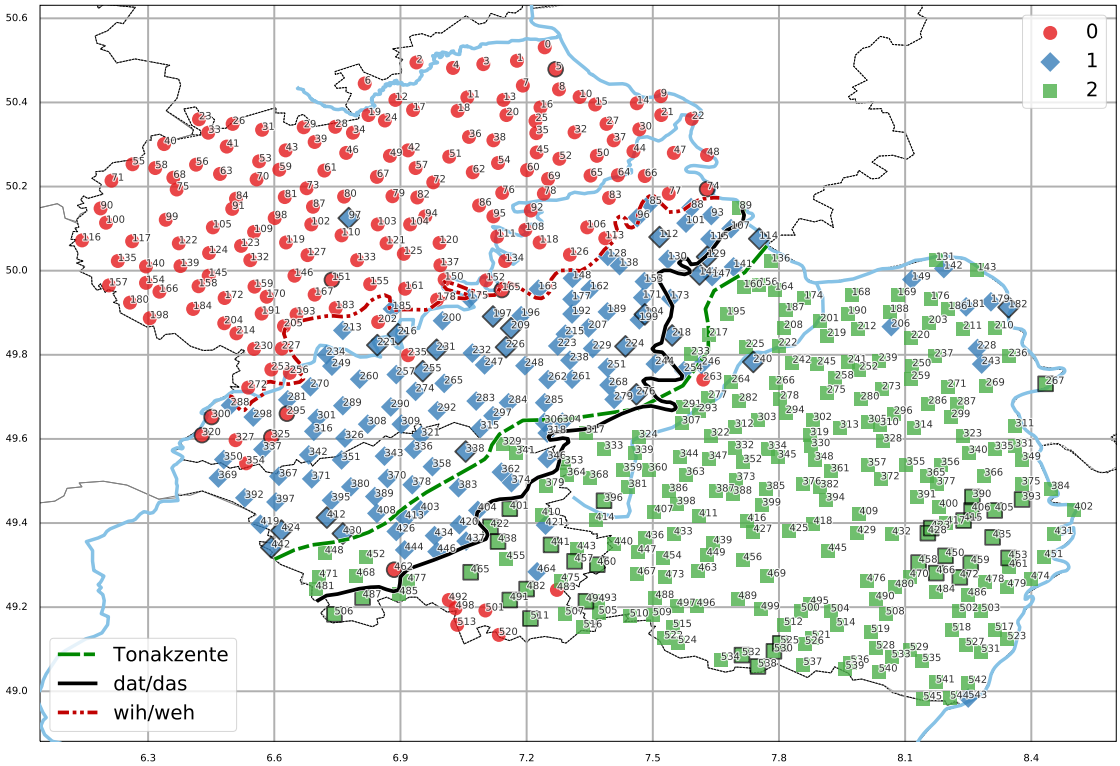

Abbildung 4.16: Mittlere Verteilung der einzelnen phonetischen Eigenschaften nach Cluster und aufgeteilt nach historischen Lautklassen und den Langvokalen des Mittelhochdeutschen für KMEANS2.

Tabelle 4.2: Die 10 höchstsignifikanten ($p$-value < 0.001) und alle nicht signifikanten ($p$-value > 0.05) Eigenschaften für verschiedene Clusterings auf dem Datenset für die historischen Langvokale des Mittelhochdeutschen.

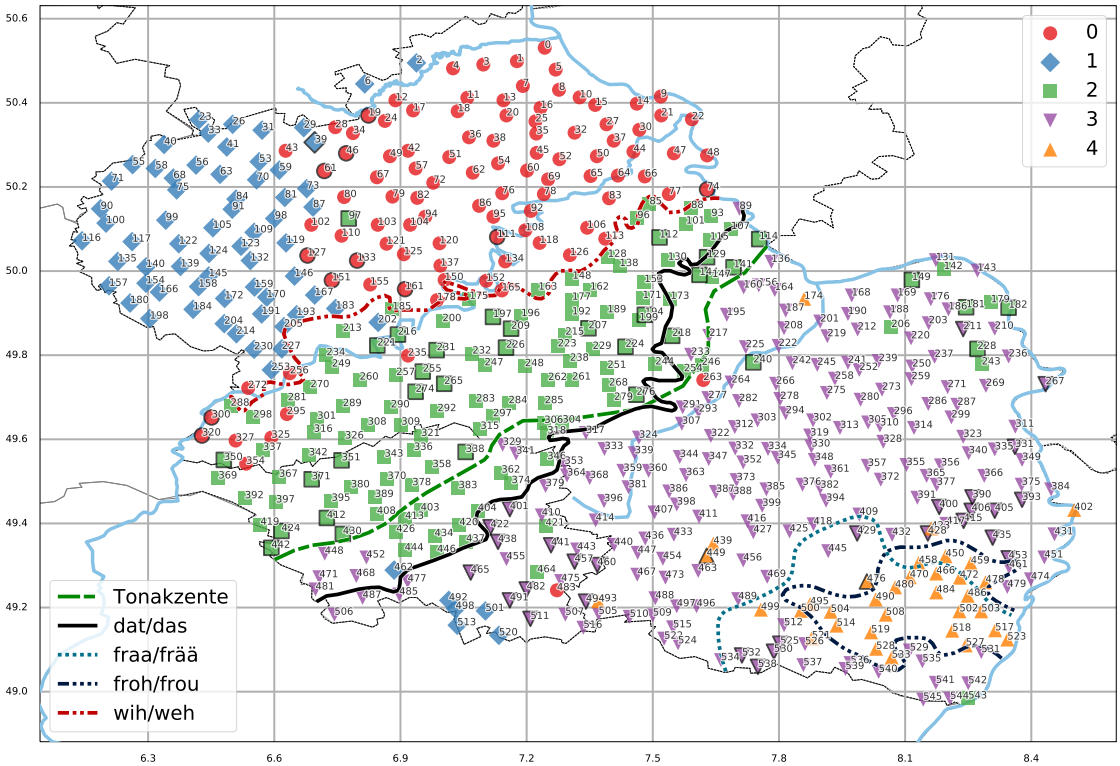
| | KMEANS2 | WARD3 | WARD5 |
|---|---|---|---|
| signifikant | LoweredClose | NearFront | NearFront |
| | NearFront | LoweredClose | LoweredClose |
| | CloseMid | CloseMid | CloseMid |
| | NearBack | TA1 | TA1 |
| | Front | TA2 | Short |
| | TA2 | NearBack | TA2 |
| | Short | Front | Front |
| | TA1 | DiphOpenCentral | NearBack |
| | DiphOpenCentral | Short | DiphLowered-CloseNearFront |
| | OpenMid | DiphLoweredClose-NearFront | DiphOpenCentral |
| nicht signifikant | Unround | Nil | Nil |
| | Long | | |
| | Nil | | |

Lachen (450) direkt nördlich des Südpfälzischen Reliktgebietes am östlichen Rand des Untersuchungsgebietes. Während das nördliche 0-Cluster weiterhin deutlich getrennt ist, ist eine saubere Trennung zwischen dem 1- und dem 2-Cluster in diesem Gebiet nicht ohne weiteres möglich. Auch finden sich gerade im nördlichen Bereich des 2-Clusters noch Orte aus dem 1-Cluster. Insgesamt ist die Stabilität dieses Clusterings mit einem Silhouettenkoeffizenten von 0.23 und einem Calinski-Harabasz-Wert von 231.50 deutlich geringer als bei einem Zweierclustering.

Das Fünferclustering in Abbildung 4.17b zeigt zwei neue Unterräume. Zum einen die Ost-West-Teilung des nördlichen Bereichs (dem 0-Cluster in KMEANS2 und WARD3) und ein neues Gebiet als 4-Cluster im Bereich des südpfälzischen Reliktgebietes. Auffällig ist, dass das Monophthonggebiet um den Ort Eschringen (501) zum 1-Cluster gezählt wird. Abbildung 4.18 liefert eine Erklärung für die generierten Cluster. Das 1-Cluster wird sehr stark von der hohen Frequenz der Eigenschaft *Short* bestimmt, die in allen Lautklassen bis auf *mhd. â* verstärkt auftritt und komplementär zu einer niedrigeren Frequenz von *Long* in den Lautklassen *mhd. æ*, *mhd. ê*, *mhd ô* und *œ* ist. Die Verteilung der einzelnen phonetischen Eigenschaften auf Seite 99 deutete bereits auf ein Gebiet mit einer hohen *Short*-Frequenz hin. Dieses




(a) WARD3




(b) WARD5

Abbildung 4.17: WARD3- (a) und WARD5-Clustering (b) für das Datenset der historischen Langvokale.

Gebiet scheint durch das 1-Cluster definiert zu sein, welches im Raum der Westeifel liegt. Zudem weist dieses Gebiet im Vergleich zu dem 0-Cluster eine hohe Frequenz von [ɪ]-Eigenschaften in *mhd. æ* auf. Historisch lässt sich dieses Phänomen an den Wenkerkarten „nähen" (WA:260)[155] und „mähen" (WA:525)[156] ausmachen. Die 2- und 3-Cluster sind sich ähnlich, was die allgemeine Verteilung der Eigenschaften angeht, unterscheiden sich aber in den Tonakzenten, die im 3-Cluster nicht auftreten und der Eigenschaft *Unround*, die im 2-Cluster eine hohe Frequenz aufweist.

Auch lässt sich die *Nasal*-Eigenschaft verorten. Das häufigste Vorkommen ist das im 3-Cluster, gefolgt von dem im 4-Cluster. Im 2-Cluster ist die Frequenz mit 0.06 so gering, dass sie als statistische Unruhe[157] abgetan werden kann. Nasalität in Langvokalen ist damit eine RHEINFRÄNKISCHE Eigenschaft.

Das 4-Cluster grenzt sich von dem 3-Cluster insofern ab, als *mhd. ô* meistens in der Form des [o͡ʊ]-Diphthongs erscheint, repräsentiert durch die Eigenschaften *DiphLoweredCloseNearBack* und *DiphOpenMidBack*, wohingegen das 3-Cluster dort [oː] als Lautvariante benutzt. Diese Verteilung wird auch bei Lauten zu *mhd. â* verwendet, allerdings in etwas abgeschwächter Form. Während die Laute zu *mhd. ê* und *mhd. œ* im 3-Cluster zu [eː] zusammenfallen, gekennzeichnet durch eine hohe Frequenz von *CloseMid* und *Front*, bleibt im 4-Cluster ein Unterschied erhalten. Die Laute zu *mhd. œ* werden weitestgehend durch den [ɛɪ̯]-Diphthong ausgedrückt, wohingegen *mhd. ê* das [eː]-Phon weitesgehend behält. Ein ähnliches Verhalten lässt sich auch für *mhd. æ* und *mhd. œ* beobachten.

Die Stabilitätsmetriken sind etwas niedriger als beim Dreierclustering mit einem Silhouettenkoeffizenten von 0.20 und einem Calinski–Harabasz-Wert von 156.33. Der ARI beträgt allerdings nur noch 0.69[158], was als Indiz gesehen werden kann, dass die Cluster nicht mehr so deutlich getrennt sind wie zum Beispiel das KMEANS2-Clustering. Ein Bootstrapping (siehe Abschnitt A.5, Abbildung A.3b auf Seite 220) zeigt eine größere Varianz in den Labelzuordnungen als in den Zweier- und Dreiclusterings, aber die zugeordneten Label sind weiterhin deutlich dominant, insbesondere bei dem 4-Cluster und dem 0-Cluster im UMLAUTGEBIET. So zeigt sich, dass die Trennung zwischen dem 0- und dem 1-Cluster nicht so stabil ist. Dies lässt das 1-Cluster als „Westeifelgebiet" eher als Unterraum denn als eigenständiges Gebiet erscheinen. Ein weiteres instabiles Gebiet findet sich zwischen der Tonakzentgrenze und der *dat/das*-Isoglosse. Dieser Bereich, der weitestgehend dem 2-Cluster zugeordnet ist, kann als ein Übergangsgebiet interpretiert werden.

### *Einfluss der Tonakzente*

Auf das Zweierclustering haben die Tonakzente offensichtlich keinen großen Einfluss (ARI von 0.93), da die Grenze deutlich woanders liegt. Für höhe-

155 <https://www.regionalsprache.de/SprachGis/RasterMap/wa/269>, abgerufen 30.01.2018.

156 <https://www.regionalsprache.de/SprachGis/RasterMap/wa/269>, abgerufen 30.01.2018.

157 Engl. Noise.

158 Aufgrund der Art, wie diese Stabilitätsmetrik berechnet wird, kann ein ARI >0.5 immer noch als gut interpretiert werden. Ein ARI von 0 bedeutet eine zufällige Labelzuordnung.

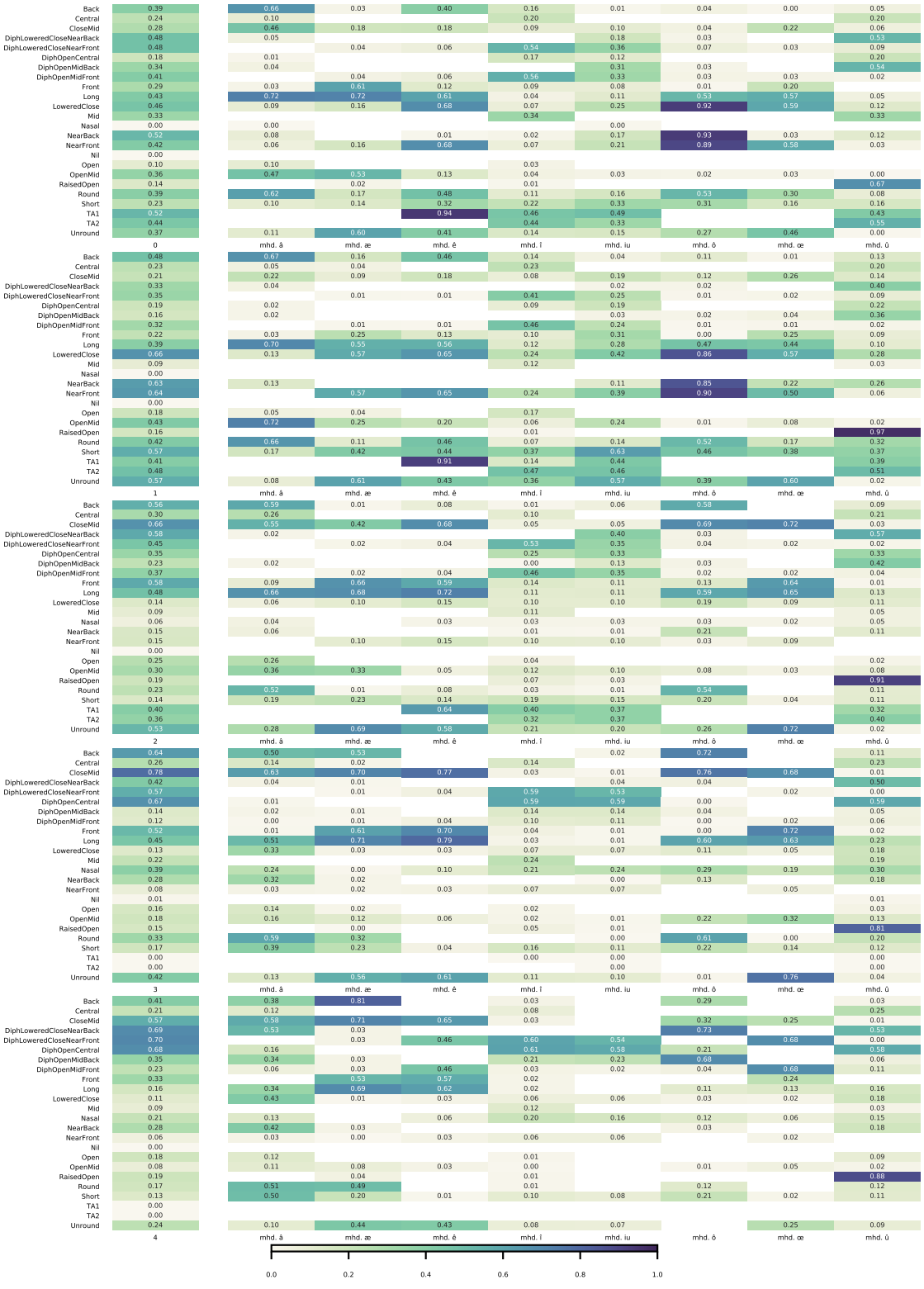



Abbildung 4.18: Mittlere Verteilung der einzelnen phonetischen Eigenschaften nach Cluster und aufgeteilt nach historischen Lautklassen und den Langvokalen des Mittelhochdeutschen für WARD5.

re Clusterings lässt doch ein merklicher Einfluss der Tonakzente feststellen. So zeigt Abbildung 4.19 das WARD5 Clustering ohne die Tonakzente. Man erkennt, dass das 0- und 1-Cluster weitestgehend stabil bleibt. Das 2-Cluster geht im südlichen Bereich allerdings über die *dat/das*-Isoglosse hinaus und umfasst noch das Gebiet des Saarlandes im RHEINFRÄNKISCHEN. Insgesamt sind die 2- bis 4-Cluster weniger homogen. Ein völlig anderes Raumbild ergibt sich jedoch nicht, dennoch können die Tonakzente für eine Stabilisierung entlang der *dat/das*-Isoglosse verantwortlich gemacht werden.

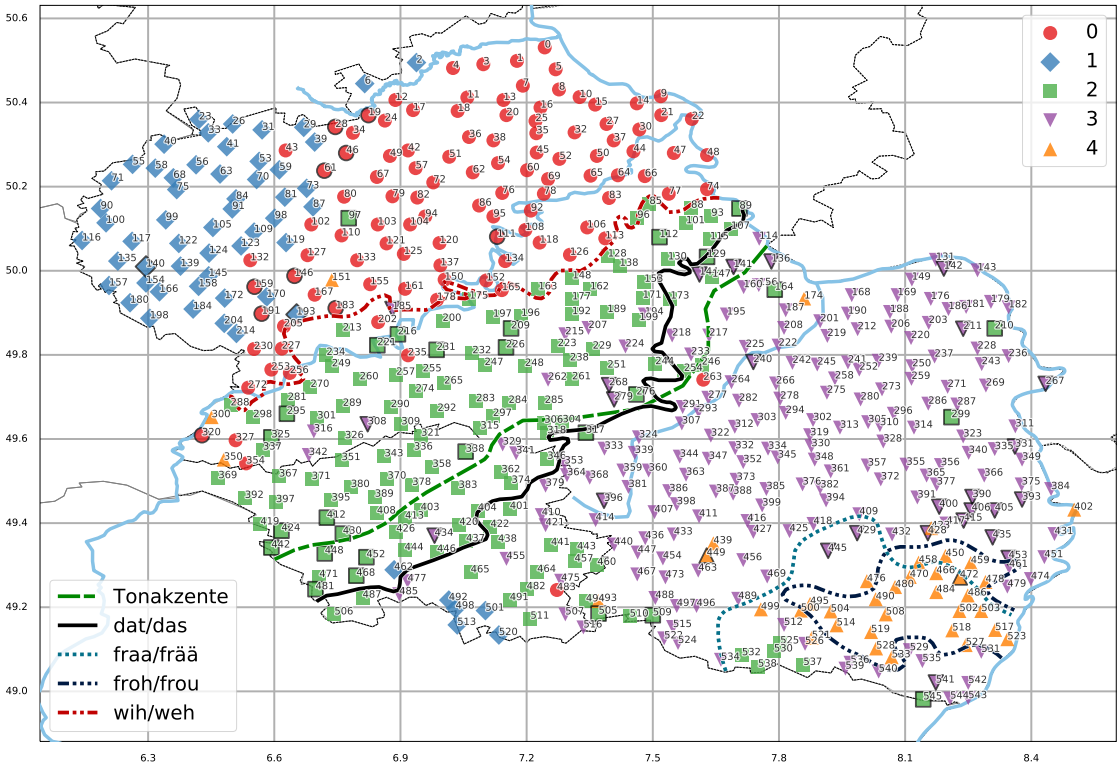



Abbildung 4.19: WARD5-Clustering für das Datenset der historischen Langvokale ohne Berücksichtigung der Tonakzente.

*Bemerkungen*

Die historischen Langvokale des Mittelhochdeutschen liefern ein gut separierbares Clustering für ein k von 2. Die Hauptgrenze verläuft in diesem Fall aber nicht entlang der Tonakzent- und *dat/das*-Grenze, sondern zeichnet sich in erster Linie durch einen Gegensatz von *LoweredClosed* und *OpenMid* bei dem Öffnungsgrad der Vokale aus. Dieser Gegensatz ist als Reihenvertauschung bekannt und wird in Schmidt (2015) ausführlich diskutiert.

Höhere Clusterings liefern eine zusätzliche Separierung an der Hauptgrenze entlang der Tonakzentgrenze und der *dat/das*-Isoglosse, die bereits aus dem ALLE-Experiment bekannt ist. Diese Grenze ist insofern interessant, als das verwendete Datenset nur auf den Eigenschaften zu den histori-

schen Langvokalen beruht, die *dat/das*-Isoglosse aber auf konsonantischen Lauteigenschaften basiert. Ein Bootstrapping zeigt aber, dass das Gebiet zwischen den beiden Grenzen instabiler ist als die übrigen Gebiete in den angrenzenden Clustern. Das Cluster (4-Cluster in WARD5), welches mit dem Kern des SÜDPFÄLZISCHEN RELIKTGEBIETES koinzidiert, zeichnet sich durch eine hohe Frequenz an Diphthongen aus. In der Westeifel bildet sich im Fünferclustering ein Gebiet heraus, das sich durch ein höheres Auftreten der *Short*-Eigenschaft auszeichnet. Allerdings zeigt ein Bootstrapping, dass sich dieses Gebiet nicht sauber von dem angrenzenden Cluster trennen lässt. Somit sollte es eher als Unterraum aufgefasst werden.

## 4.4 UNTERSUCHUNG DER OBSERVATIONEN ZU DEN LAUTEN DER MITTELHOCHDEUTSCHEN KURZVOKALE

*Vorverarbeitung*

Das Experiment zu den historischen Kurzvokalen (KURZ) basiert auf 251037 inferierten phonetischen Eigenschaften, verteilt auf die 546 Orte des Untersuchungsgebietes. Die Datenumwandlung erfolgt dabei wie bei dem LANG-Experiment. Die Verteilung der einzelnen Eigenschaften ist in Abbildung 4.20 dargestellt. Auffällig sind hierbei die vielen negativen Ausreißer bei *Unround* und die positiven Ausreißer bei *Long*. Die vielen negativen Ausreißer zu *Unround* weisen höchstwahrscheinlich auf das UMLAUTGEBIET im Norden des Untersuchungsgebietes hin, das sich bisher immer durch eine deutliche *Round*–*Unround*-Differenz ausgezeichnet hat.

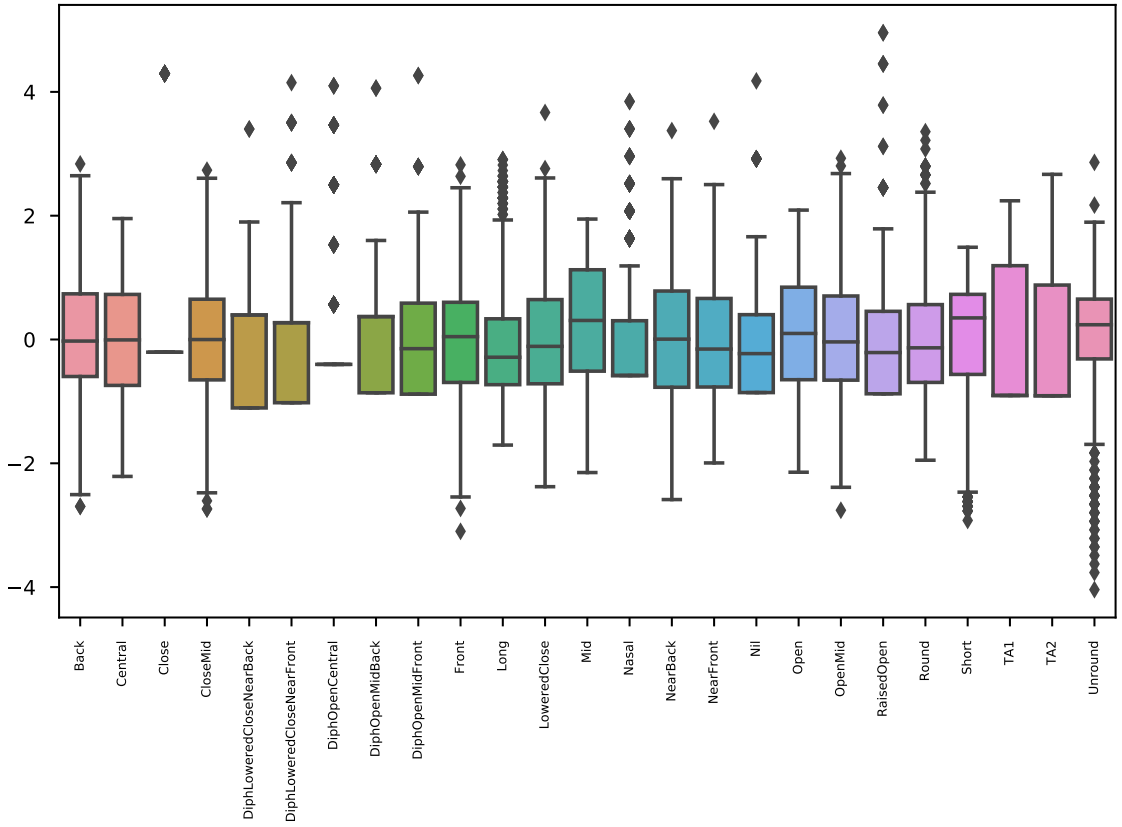



Abbildung 4.20: Verteilung der phonetischen Eigenschaften zu den Observationen der historischen Kurzvokale des Mittelhochdeutschen.

Auch findet sich wieder die konsonantische Eigenschaft *Nasal* mit einer ähnlichen Verteilung, wie im LANG-Experiment. Dies lässt vermuten, dass diese Eigenschaft in derselben Region (dem RHEINFRÄNKISCHEN) vorkommt.

Die vielen Ausreißer zu *Long* lassen sich über die Tonakzente erklären, die zu einer Dehnung des Vokals führen. Bei der Untersuchung der Langvokale (Abschnitt 4.3) hat sich bei dem Fünferclustering (siehe Seite 108) ein Gebiet

im Bereich der Westeifel mit einer hohen Frequenz an *Short*-Eigenschaften herausgebildet. Es ist zu vermuten, dass die vielen Ausreißer zu *Long* in diesem Gebiet zu finden sind.

Diese Behauptungen werden durch die Korrelationsmatrix (siehe Abbildung 4.21) gestützt. Man sieht deutlich eine hohe Korrelation zwischen den Tonakzenten und *Long* sowie eine Antikorrelation zwischen den Tonakzenten und *Nasal*. Neben der erwarteten hohen negativen Korrelation zwischen *Unround* und *Round* sind auch *Long* und *Short* deutlich negativ korreliert, was auf eine Ersetzung des Merkmals und nicht auf eine zusätzliche Eigenschaft schließen lässt. Wegen der Korrelation mit den Tonakzenten können wir inferieren, dass die *Long*-Eigenschaft stark im Gebiet des MOSELFRÄNKISCHEN vertreten ist. Ähnlich wie bei den Langvokalen gibt es weiterhin eine Antikorrelation zwischen den [ɪ] und [e] definierenden Eigenschaften.

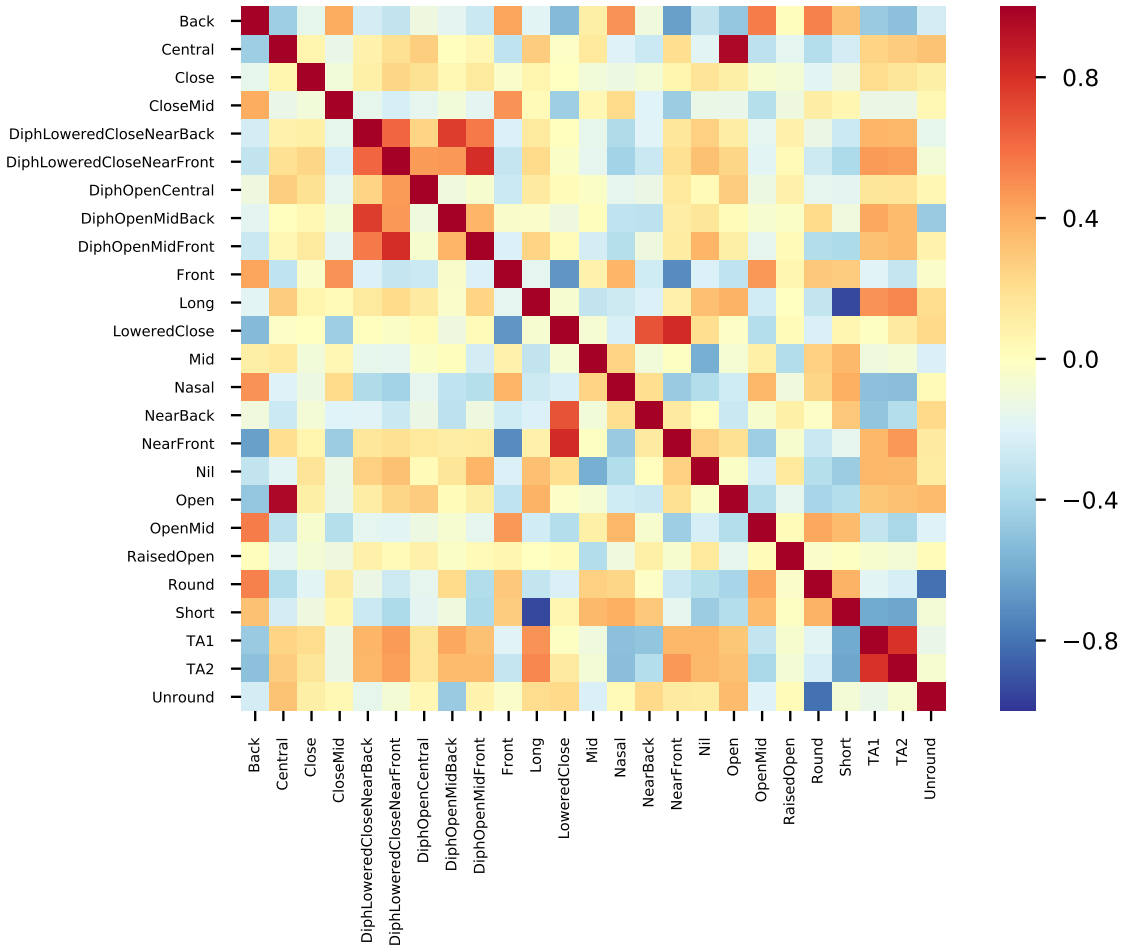



Abbildung 4.21: Korrelationsmatrix der phonetischen Eigenschaften zu den Observationen der historischen Kurzvokale des Mittelhochdeutschen.

Die Hauptkomponentenanalyse reduziert die 47 ursprünglichen Dimensionen auf 18, wobei die erste Dimension 28% der Varianz erfasst und die ersten drei Dimensionen zusammen 53%. Dies ist etwas weniger als bei dem

Langvokaldatenset, was auf eine breitere Streuung der Daten hinweist. Der Einfluss der ursprünglichen Features auf diese Dimensionen ist in Abbildung 4.22 zu finden. Die wichtigsten Features sind die Tonakzente (*TA1*, *TA2*), *Back* und *Short*.

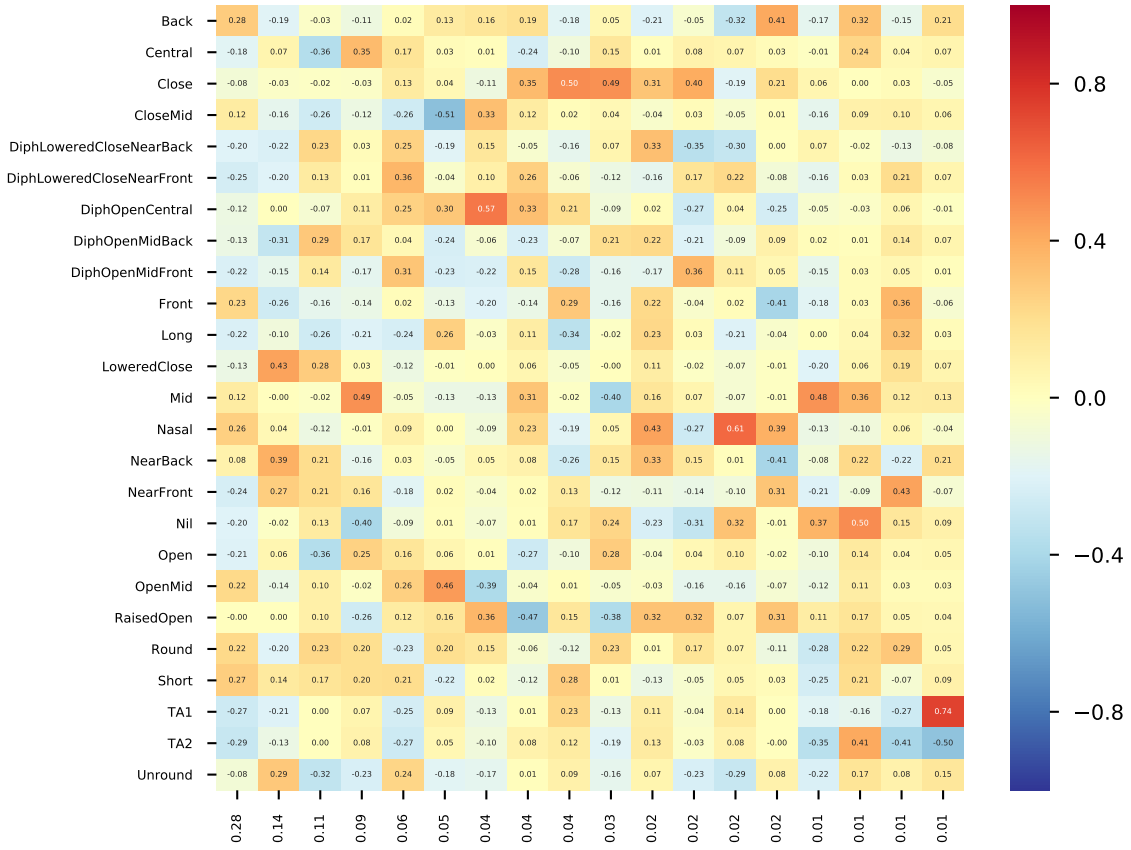



Abbildung 4.22: Anteile der Varianz der ursprünglichen Dimensionen des Kurzvokaldatensets auf die Varianz der neuen, reduzierten Dimensionen nach einer Hauptkomponentenanalyse.

Die Visualisierung der ersten drei Dimensionen des durch die PCA (Abbildung 4.23) reduzierten Datensets zeigt eine grobe Dreiteilung des Untersuchungsgebietes. Im Norden wird mit einem eher türkisen Bereich das Umlautgebiet hervorgehoben. Der Rest des Moselfränkischen wird in grün bis grünbraun dargestellt. Das Rheinfränkische erscheint weitestgehend in einem blauviolet, es findet sich aber im Bereich des südpfälzischen Reliktgebietes ein kleiner Bereich, der ähnlich wie das Umlautgebiet türkis erscheint.

*Clusteranalyse*

Das Clustering der Observationen zu den historischen Kurzvokalen des Mittelhochdeutschen ist instabiler als die bisherigen beiden Experimente. Auf den ersten Blick scheint sich eine Struktur ähnlich denen des Dreierclusterings im ALLE-Experiment abzuzeichnen, bei einer genauen Betrachtung

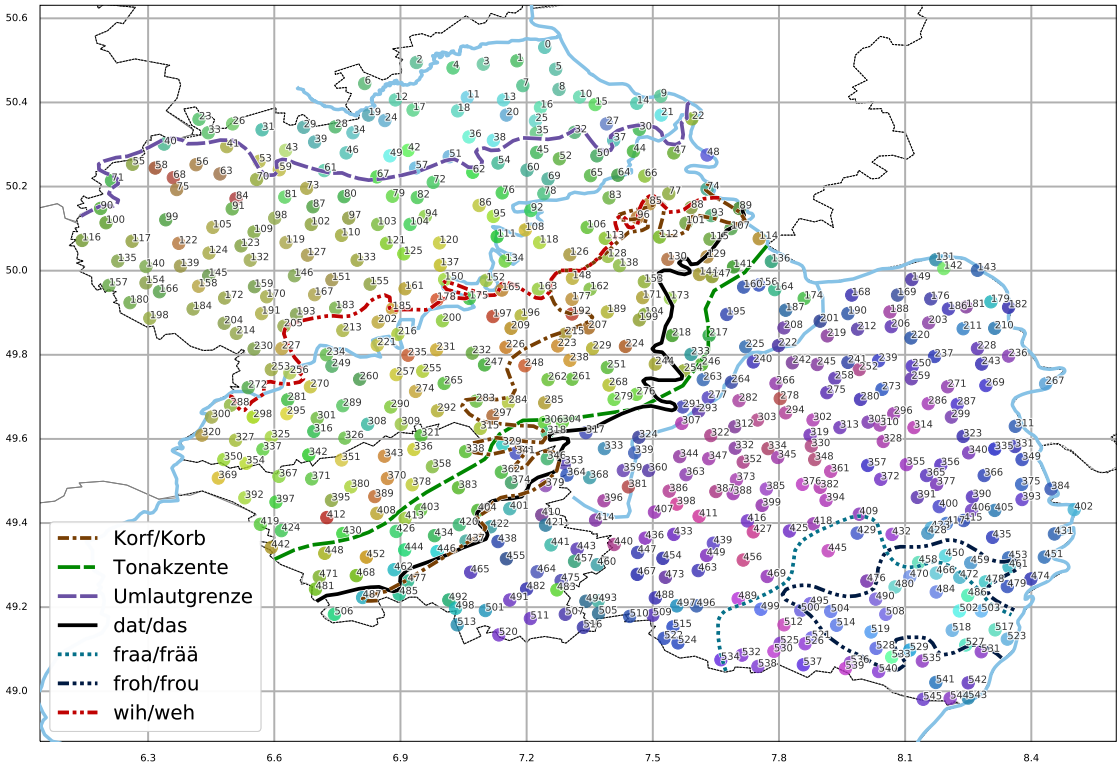
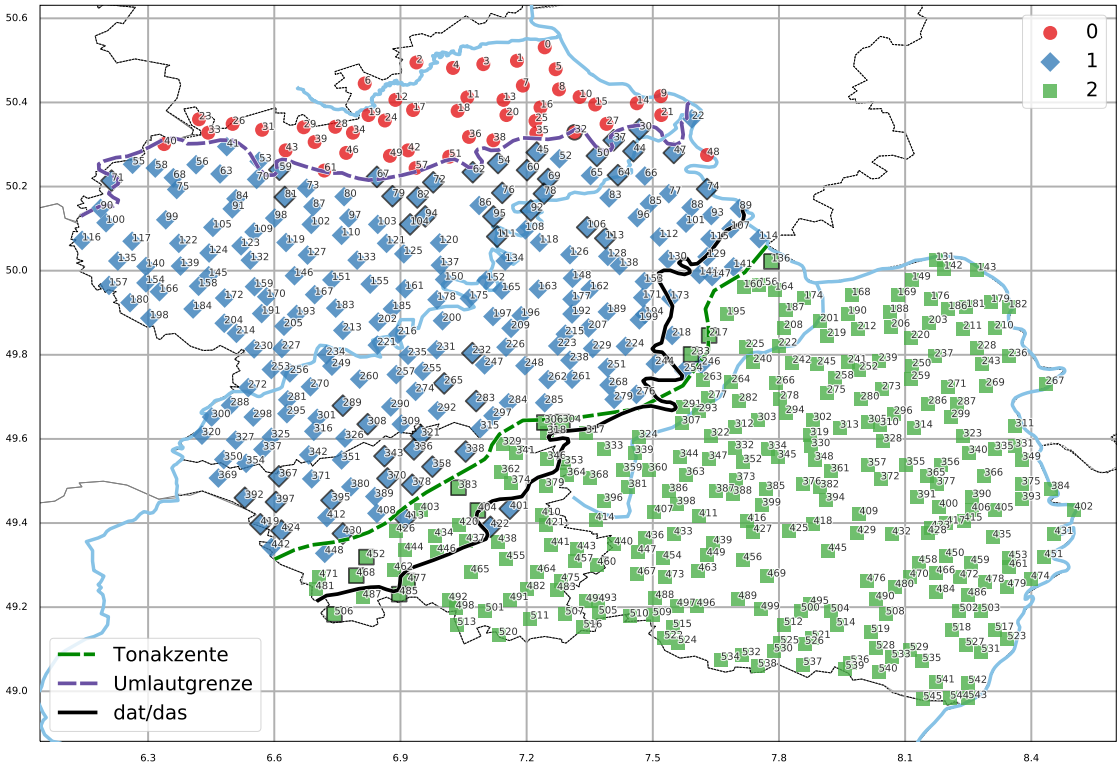
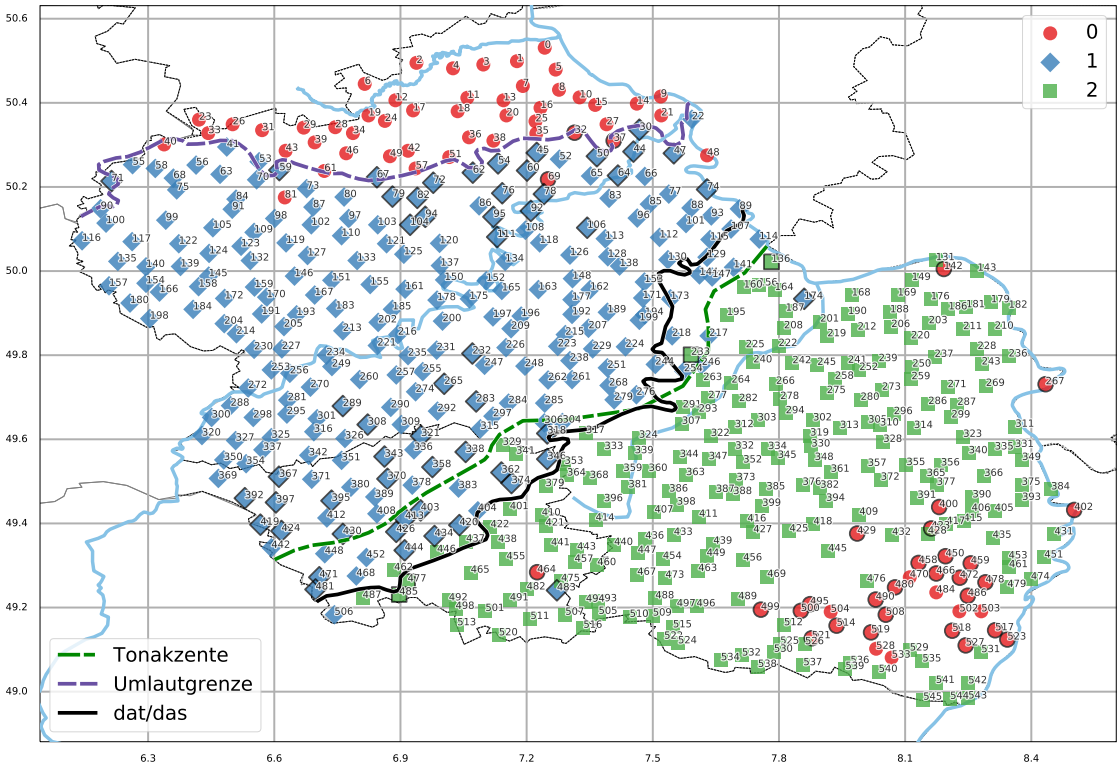



Abbildung 4.23: Räumliche Visualisierung des Kurzvokaldatensets durch die ersten drei Dimensionen einer PCA, eingefärbt nach dem HSV Farbmodell.

treten aber ein paar Eigenheiten hervor. Abbildung 4.24a zeigt das Dreierclustering nach GMM3. Es folgt der bereits in der PCA-Visualisierung vermuteten Dreiteilung des Untersuchungsgebietes. Die Grenze zwischen dem 1- und dem 2-Cluster ist aber anders als beim Datenset über alle Lauteigenschaften und dem Dreierclustering zu LANG nicht eine Mischung aus der Tonakzentgrenze und der *dat/das*-Isoglosse, sondern folgt hier deutlich stärker der Tonakzentgrenze. Das Dreierclustering nach KMEANS3 (Abbildung 4.24b) zeigt ein etwas anderes Bild. Im südlichen Untersuchungsgebiet findet sich ein Raum, der auch dem 0-Cluster, das seine Hauptausdehnung ansonsten im Norden im UMLAUTGEBIET hat, zugeordnet wird. Dieser Raum koinzidiert mit dem SÜDPFÄLZISCHEN RELIKTGEBIET. Dass zwischen diesen beiden Gebieten eine gewisse Nähe besteht, hat sich bereits in der PCA-Visualisierung angedeutet. Das heißt aber auch, dass zum einen die Form der Cluster deutlicher durch die Wahl des Algorithmus beeinflusst wird und dass die Räume weniger homogen sind. Man sieht allerdings auch, dass die meisten Orte im 0-Cluster im Süden eine negative Silhouette haben. Dies lässt vermuten, dass die Zuordnung dieser Orte zum 0-Cluster nicht völlig eindeutig ist und dieses Clustering instabiler ist als das GMM3. Ebenfalls sieht man in beiden Clusterings viele Orte mit negativen Silhouetten im 1-Cluster an dem östlichen Randgebiet zu dem 0-Cluster. Dies weist auch auf eine eher instabile Unterregion hin. Eine Nearest-Neighbor-Analyse auf den Distan-




(a) GMM3




(b) KMEANS3

Abbildung 4.24: GMM3- (a) und KMEANS3-Clustering (b) für das Datenset der historischen Kurzvokale.

zen im Datenraum zeigt, dass die Orte, die im 1-Cluster eine negative Silhouette haben, Orte aus 2-Cluster als nächsten Nachbarn haben. Nur ein paar Orte verweisen auf Orte im SÜDPFLÄZISCHEN RELIKTGEBIET. Die Metriken geben einen Silhouettenkoeffizienten von 0.23 und einen Calinski–Harabasz-Wert von 151.96 für GMM3 an und 0.17 und 144.80 für KMEANS3. Dies lässt das GMM3-Clustering etwas stabiler erscheinen. Interessanterweise liefert der ARI ein umgekehrtes Bild. Mit 0.93 ist er für KMEANS3 höher als für GMM3 mit 0.85. Dies kann als Hinweis gesehen werden, dass die historischen Kurzvokale einem weniger klaren Muster folgen als die Langvokale. Ein Bootstrapping zu GMM3 (ohne Abbildung) rekonstruiert die Ausgangscluster fast zu 100%, während das Bootstrapping zu KMEANS3 (ohne Abbildung) einen deutlichen Einfluss des 2-Clusters auf das 0-Clustergebiet im Bereich des SÜDPFÄLZISCHENN RELIKTGEBIETES aufweist.

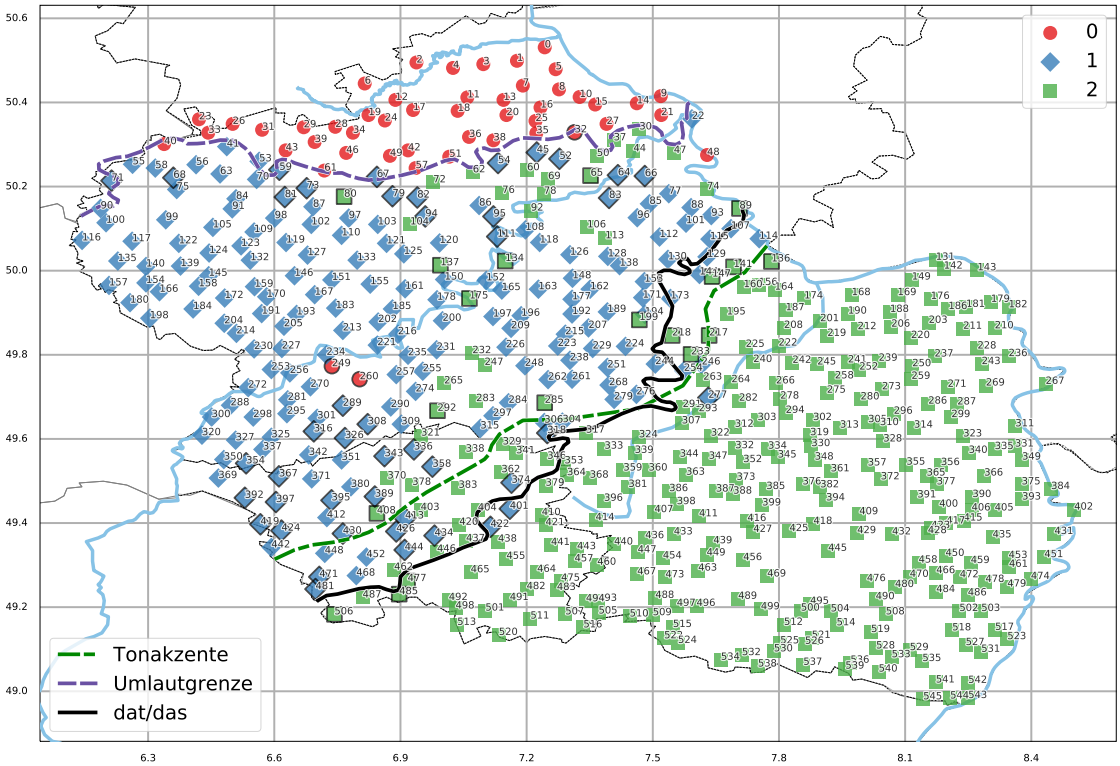
*Bedeutung der Tonakzente*

Da die Tonakzentgrenze diesmal als die bestimmende Grenze in GMM3 dient, ist eine Analyse der Daten ohne ihren direkten Einfluss von besonderem Interesse. In Abbildung 4.25a sieht man das GMM3-Clustering ohne die Tonakzente. Man erkennt, dass das Clustering ohne die Tonakzente die Hauptgrenze zwischen dem 1- und 2-Cluster sich der Form aus dem ALLE-Experiment annähert, allerdings gibt es sehr viele „Ausreißer" aus dem 2-Cluster im Gebiet des östlichen MOSELFRÄNKISCHEN. Ein Bootstrapping zu dem GMM3 zeigt einige interessante Auffälligkeiten. Zum einen wird das Hauptgebiet des 1-Clusters deutlich homogener rekonstruiert als das entsprechende Clustering, und zum anderen findet sich wieder ein 0-Cluster-Einfluss im Gebiet des SÜDPFÄLZISCHEN RELIKTGEBIETES. Die Hauptgrenze zwischen den das 1-Cluster dominierenden Orten und den das 2-Cluster dominierenden Orten verschiebt sich sogar noch deutlicher in Richtung der aus dem ALLE-Experiment bestimmten Hauptgrenze.

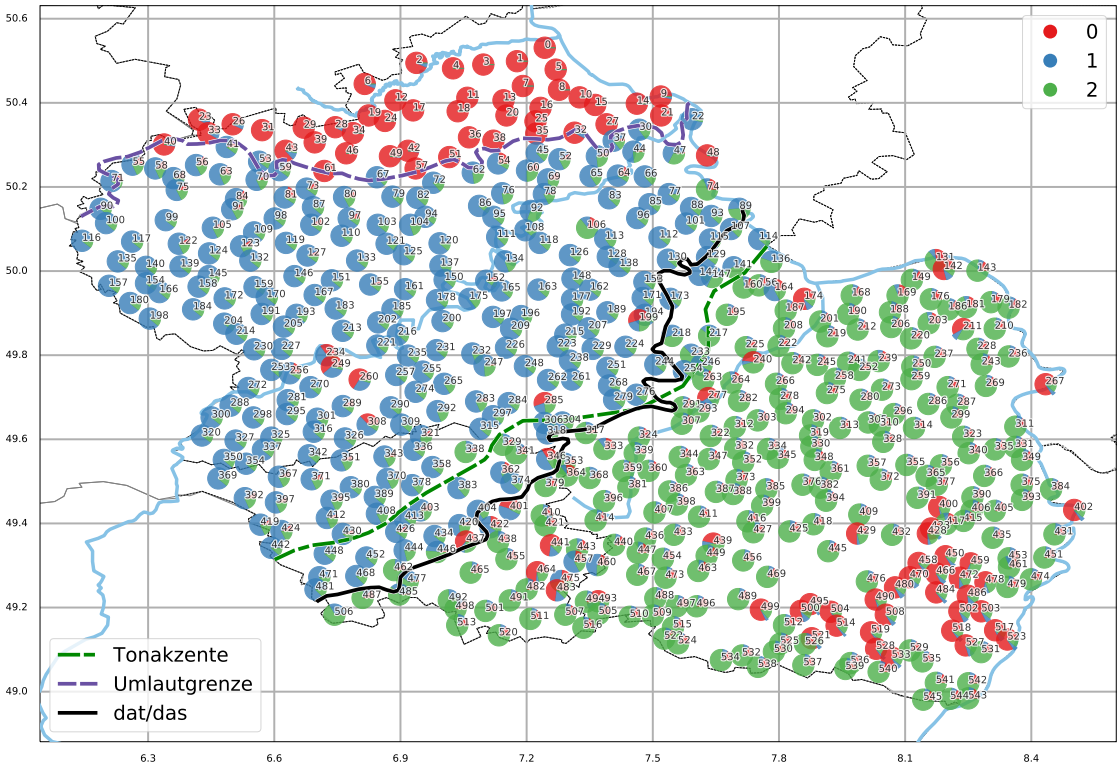
Für die historischen Kurzvokale bilden die beiden Eigenschaften *TA1* und *TA2* einen wiederum stabilisierenden Faktor, allerdings bleibt die generelle Form der Cluster erhalten. Der Silhouettenkoeffizient für das GMM3 ohne die Tonakzente beträgt immer noch 0.22 und der Calinski–Harabasz-Wert 140.30. Der Adjusted-Rand-Index ist mit 0.47 jedoch sehr niedrig, was ein Hinweis dafür sein kann, dass Zufall einen merkbaren Einfluss auf dieses Clustering hat. Da das 2-Cluster im GMM3-Clustring viele „Ausreißer" im Gebiet des MOSELFRÄNKISCHEN hat, die aber beim Bootstrapping deutlich weniger stark vertreten sind, ist anzunehmen, dass diese Datenpunkte für den niedrigen ARI verantwortlich sind. Sie können daher als Rauschen betrachtet werden. Dies spricht aber wieder für die Behauptung, dass die historischen Kurzvokale eine höhere Varianz über die gesamte Region verteilt aufweisen.

*Merkmaleinfluss*

Bei einer Betrachtung des Einflusses einzelner Features auf verschiedene Clusterings sieht man den deutlichen Einfluss der Tonakzente. Tabelle 4.3 zeigt die zehn einflussreichsten Features in verschiedenen Clusterings. Für




(a) GMM3 ohne Tonakzente




(b) Boostrapping zu GMM3

Abbildung 4.25: GMM3-Clustering (a) und Boostrapping zu GMM3 (b) für das Datenset der historischen Kurzvokale ohne Tonakzente.

GMM3 haben wir drei Hauptgegensätze, die das Clustering bestimmen. Als wichtigstes Merkmal stechen eindeutig die Tonakzente hervor, dann der *Round–Unround*-Gegensatz und der *Long–Short*-Gegensatz. Bei KMEANS3 schiebt sich mit *DiphOpenMidBack* eine Diphthongeigenschaft an die erste Position vor die Tonakzente und *Round*. Auch sind noch weitere Diphthongeigenschaften hoch gewichtet. Da der Hauptunterschied zu GMM3 das Herausstellen des SÜDPFÄLZISCHEN RELIKTGEBIETES als Teil des 0-Clusters ist, kommt die Vermutung auf, dass der [ɔ͡ʊ]-Diphthong als Kombination der Eigenschaften *DiphOpenMidBack* und *DiphLoweredCloseNearBack* eine Gemeinsamkeit zwischen dem UMLAUTGEBIET und dem SÜDPFÄLZISCHEN RELIKTGEBIET bildet.

Tabelle 4.3: Die zehn höchstsignifikanten ($p$-value < 0.001) Eigenschaften für verschiedene Clusterings auf dem Datenset für die historischen Kurzvokalen des Mittelhochdeutschen. Das einzige als nicht signifikant klassifizierte Feature ist *RaisedOpen* für KMEANS3

| GMM3 | KMEANS3 | KMEANS5 |
|---|---|---|
| TA2 | DiphOpenMidBack | TA1 |
| TA1 | TA2 | Round |
| Round | Round | TA2 |
| Unround | TA2 | DiphOpenMidBack |
| Short | DiphLowered-CloseNearBack | Unround |
| Long | Unround | DiphLowered-CloseNearBack |
| DiphOpenMidBack | Short | Short |
| Nasal | DiphLoweredClose-NearFront | NearFront |
| Back | Nasal | LoweredClose |
| NearBack | Back | DiphOpenMidFront |

*Höhere Clusterings*

Das Fünferclustering in Abbildung 4.26 zeigt nun alle angesprochenen Räume. Der Silhouettenkoeffizient beträgt allerdings nur noch 0.14 und der Calinski–Harabasz-Wert beträgt 124.51. Der ARI ist mit 0.90 aber ungewöhnlich hoch. Auffällig sind dabei die vielen 3-Cluster-Orte mit negativen Silhouetten, die sich besonders entlang der Entrundungsgrenze finden und die vereinzelten 4-Cluster-Orte in der Nähe der *dat/das*-Isoglosse. Abbildung 4.27 zeigt die mittlere Verteilung der Cluster. Hier sieht man auch die Verbindung zwischen dem 0-Cluster und dem 4-Cluster, deren Regionen bei KMEANS3 als ein Cluster zusammengefasst wurden und auch beim Bootstrapping von GMM3 ohne Berücksichtigung der Tonakzente Ähnlichkeiten aufweisen. Die

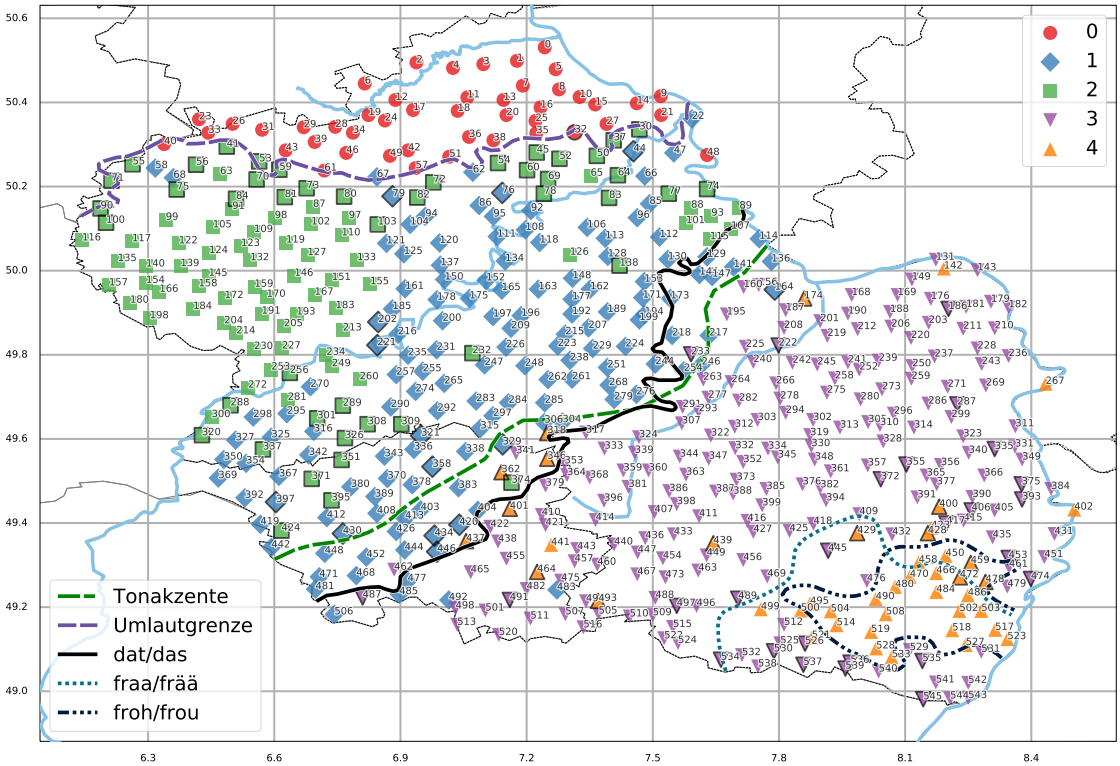



Abbildung 4.26: KMEANS5-Clustering auf dem Kurzvokaldatenset.

Hauptgemeinsamkeit ist die *DiphOpenMidBack*-Eigenschaft, die sich besonders in dem [ɔ͡ʊ] manifestiert. Der vermeintliche Zusammenhang der Regionen, die mit dem UMLAUTGEBIET und dem SÜDPFÄLZISCHEN RELIKTGEBIET, kann damit wohl als ein künstlicher angesehen werden; ein echter struktureller Zusammenhang ist nicht gegeben. Die Orte im SÜDPFÄLZISCHEN RELIKTGEBIET wurden dem 0-Cluster zugeordnet, weil sie bei einem k von 3 noch nicht so stark sind, dass sie als eigenständiges Cluster hervortreten können und sind deswegen dem nächstbesten Cluster zugeordnet. Ansonsten gibt es wenig Überraschungen. Der *Round–Unround*-Gegensatz ist weiter sehr dominant im 0-Cluster. Der hohe Wert zu *Mid* ist mit Vorsicht zu betrachten, da dies eine Eigenschaft ist, die nur sporadisch auftritt. Da die Eigenschaften skaliert sind, neigen seltene Eigenschaften zu stärkeren Sprüngen in der Varianz. Im 1-Cluster treten die [ɪ] und [ʊ] erzeugenden Eigenschaften häufiger auf als die Eigenschaften zu [e] und [o]. Ob dies ein Hinweis auf eine Reihenvertauschung ist, lässt sich nicht sagen, da die Raumstrukturen andere sind, aber die Gesamtverteilung lässt auf eine weit weniger homogene Struktur schließen. Dies lässt vermuten, dass eine konsequente Reihenvertauschung wie in den Langvokalen nicht stattgefunden hat, sondern sich die Lautvarianten über ein größeres Spektrum verteilen. Im 2-Cluster, welches die Westeifel markiert, sind die [e] definierenden Eigenschaften häufiger als die für [ɪ], zudem taucht die *Nil*-Eigenschaft auf, die den Vokalausfall markiert. Außerdem ist, wie bereits vermutet, die *Long-*

Eigenschaft überfrequent auf Kosten der *Short*-Eigenschaft. Im 3-Cluster tritt die konsonantische *Nasal*-Eigenschaft auf, ansonsten ist dieses Cluster durch den Mangel an Diphthongeigenschaften definiert. Im 4-Cluster sind diese Eigenschaften hingegen sehr häufig.

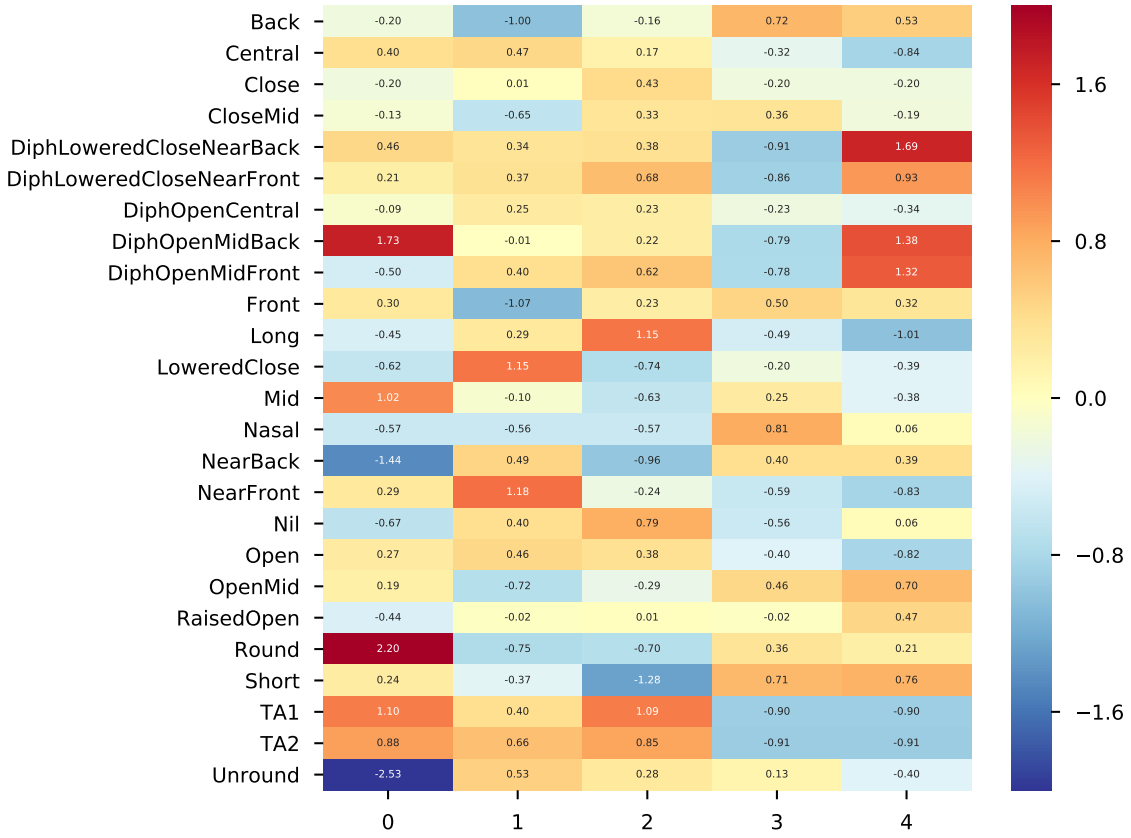



Abbildung 4.27: Mittlere Verteilung der einzelnen phonetischen Eigenschaften nach Cluster für KMEANS5 und nach den historischen Kurzvokalen des Mittelhochdeutschen.

Die Auflistung der Merkmalverteilung nach den einzelnen Lautklassen ist in Abbildung 4.28 dargestellt. Das 0-Cluster wird von der *Round*-Eigenschaft dominiert, besonders bei *mhd. ö* und *mhd. ü*, da diese Klassen fast ausschließlich durch [ʏ], [ø] und bei *mhd. ö* zusätzlich noch durch [œ] realisiert werden.

Das 2-Cluster hat sehr viele Ausprägungen mit einem mittleren Wert. Die vielen Orte mit negativen Silhouettenkoeffizienten überschatten bei diesem Cluster einzelne Ausprägungen, was zu einem Cluster ohne extreme Werte führt. Das 1-Cluster verhält sich bis auf die Rundung ähnlich. Auch hier ist eine Reihendrehung als dominierendes Merkmal für eine Lautklasse nicht auszumachen. Das 2-Cluster zeigt nun sehr hohe Werte für *Long* und niedrige für *Short* für die Westeifel. Auffällig ist zudem noch die Realisierung von *mhd. i* als [a] auf Basis von *Central* und *Open*. Dies ist allerdings kein allgemeines Phänomen für das 2-Cluster, sondern basiert auf einzelnen Wörtern

wie zum Beispiel „gebissen“[159]. Insgesamt verfügen das 1- und 2-Cluster über ein deutlich breiteres Spektrum an Lautrealisierungen in den einzelnen Lautklassen. Das 3-Cluster hat *Short* als sehr dominante Eigenschaft und zeigt eine deutliche Abwesenheit an Diphthongeigenschaften, ansonsten folgt es einer ähnlichen Verteilung wie das 1-Cluster. Auffällig ist zudem die hohe Frequenz von *Back* und *CloseMid* in *mhd. e/ä* und *mhd. o*. Das 4-Cluster unterscheidet sich von dem 3-Cluster durch das zusätzliche Auftreten von Diphthongen bei *mhd. e/ä*, *mhd. o* und *mhd. ö*.

### BEMERKUNGEN

Das Clustering der historischen Kurzvokale ähnelt auf den ersten Blick dem Clustering zu ALLE. Ein Dreiercluster unterteilt den Raum in das Moselfränkische mit dem nördlichen Umlautgebiet und das Rheinfränkische im Süden. Die Hauptgrenze wird dabei sehr stark von den Tonakzenten beeinflusst. Eine Ausklammerung der Tonakzente führt zu einer leichten Anpassung der Hauptgrenze an die bereits aus dem ALLE-Experiment bekannte Grenze, allerdings auf Kosten der Clusterstabilität und Konnektivität.

Insgesamt ist ein Clustering auf dem Kurzvokaldatenset deutlich instabiler als auf dem Langvokaldatenset und sogar auf dem Datenset über alle Vokale. Die Wahl des Clusteralgorithmus hat außerdem Einfluss darauf, in welcher Reihenfolge einzelne Cluster gebildet werden. Bei niedrigen k werden häufig das Umlautgebiet und das südpfälzische Reliktgebiet zu einem Cluster zusammengefasst, da beide Gebiete eine hohe Frequenz von *DiphOpenMidBack* haben.

Höhere Cluster weisen auf weitere Strukturen hin, wie das *Long*-Gebiet in der Westeifel. Dieses Gebiet zeigt sich beim Clustering der Langvokale mit einer dominanten *Short*-Eigenschaft. Dies lässt vermuten, dass in diesem Gebiet die Lang- und Kurzvokale teilweise vertauscht sind.

Eine Reihenvertauschung wie in den Langvokalen ist in diesen Clustern nicht so deutlich zu bemerken. Dies kann allerdings daran liegen, dass sich keine Cluster entlang der *wih/weh*-Isoglosse bilden, die als Grenze der Reihenvertauschung gesehen werden kann. Dies spricht wiederum dafür, dass die Reihenvertauschung in den Kurzvokalen nicht in demselben Maß stattfand wie bei den Langvokalen, weil sich ansonsten diese Isoglosse auch in den Clusterings wiederfinden sollte. Im Umlautgebiet gibt sich eine Tendenz, *mhd. i* mit den Eigenschaften *CloseMid* und *Front* zu realisieren. Jedoch verhält sich *mhd. ë* komplementär dazu. Insgesamt ist die Realisierung der Laute der historischen Lautklassen der Kurzvokale deutlich variantenreicher als in den Langvokalen. Dies trifft besonders auf das Moselfränkische zu.

159 <https://www.regionalsprache.de/SprachGis/VectorMap/mrhsa/3/182>, abgerufen 30.01.2018.

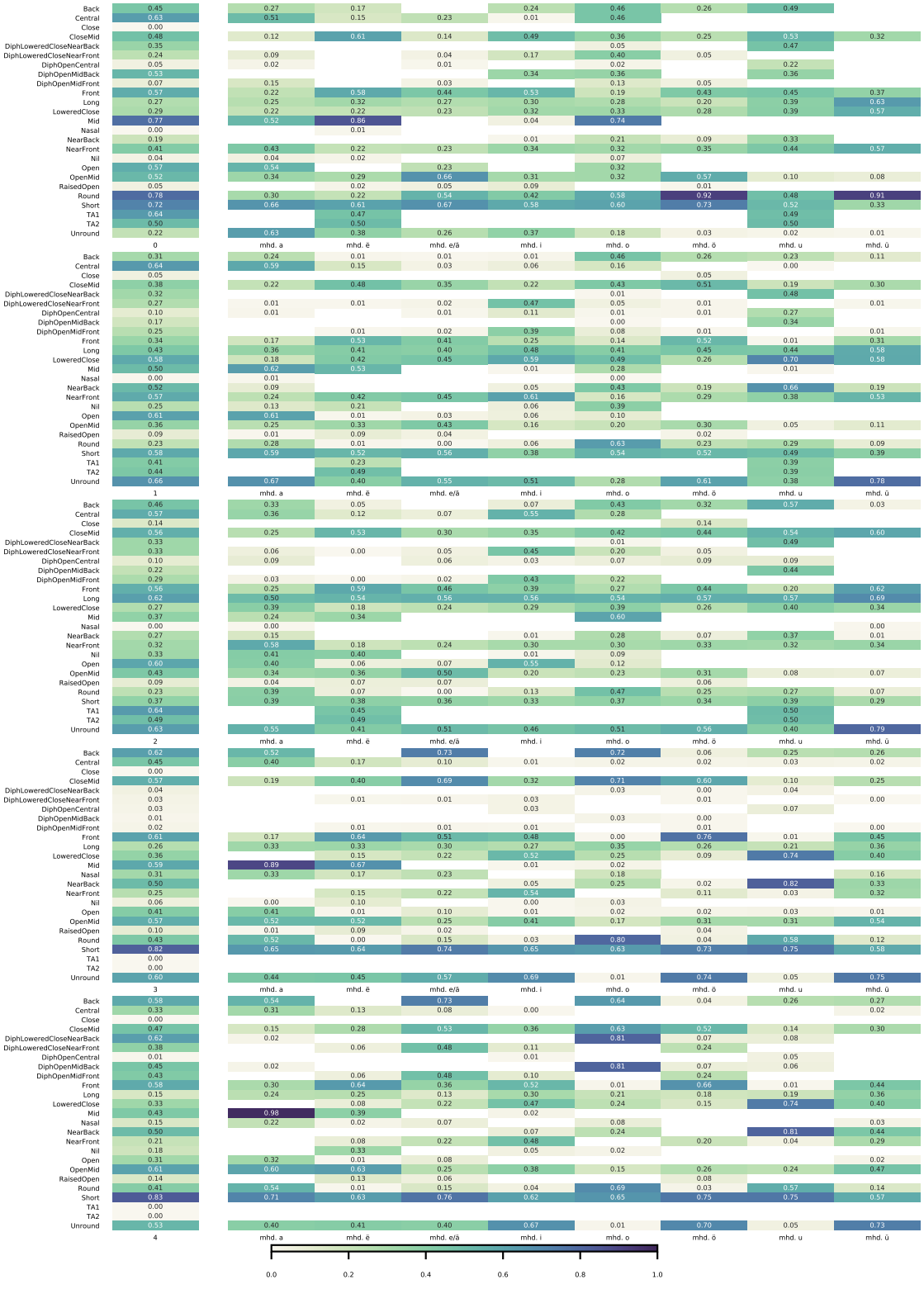



Abbildung 4.28: Mittlere Verteilung der einzelnen phonetischen Eigenschaften nach Cluster und aufgeteilt nach historischen Lautklassen und den Kurzvokalen des Mittelhochdeutschen für KMEANS5.

## 4.5 UNTERSUCHUNG DER OBSERVATIONEN ZU DEN LAUTEN DER HISTORISCHEN KLASSE DER WESTGERMANISCHEN KONSONANTEN

*Vorverarbeitung*

Die vierte Untersuchung (WG) basiert auf dem Datenset für die historischen Konsonanten des Westgermanischen und besteht aus 234796 inferierten phonetischen Eigenschaften, die in das (546, 47)-dimensionale Datenset umgewandelt werden. Abbildung 4.29 zeigt die Verteilung der phonetischen Eigenschaften in diesem Datenset.

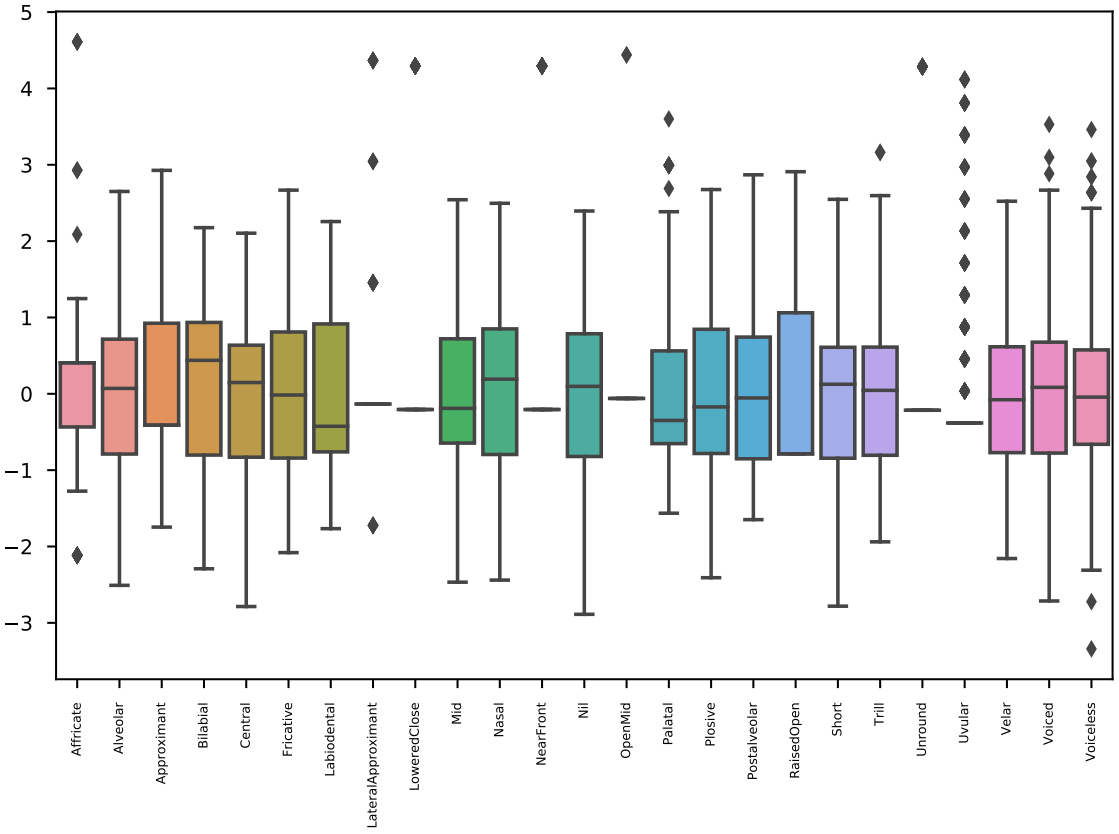



Abbildung 4.29: Verteilung der phonetischen Eigenschaften zu den Observationen der westgermanischen Konsonanten.

Neben den konsonantischen Eigenschaften sind mit *Central*, *LoweredClose*, *Mid*, *NearFront*, *RaisedOpen*, *Short* und *Unround* auch einige vokalische vertreten. Dabei gibt es zwei unterschiedliche Verteilungsmuster. Während *Central*, *Mid*, *Short* und *RaisedOpen* ein Verteilungsspektrum haben, ist die Verteilung von *LoweredClose*, *NearFront*, *OpenMid* und *Unround* bei o bis auf Ausreißer. Daraus kann man schließen, dass sich diese Eigenschaften nur auf ein sehr isoliertes Phänomen beziehen, während die anderen verbreiteter sind. Ansonsten gibt es in diesem Datenset deutlich weniger auffällige

Ausreißer als bei dem vokalischen Datenset. Einzig *Uvular* hat übermäßig viele Ausreißer, allerdings ansonsten kein sichtbares Spektrum.

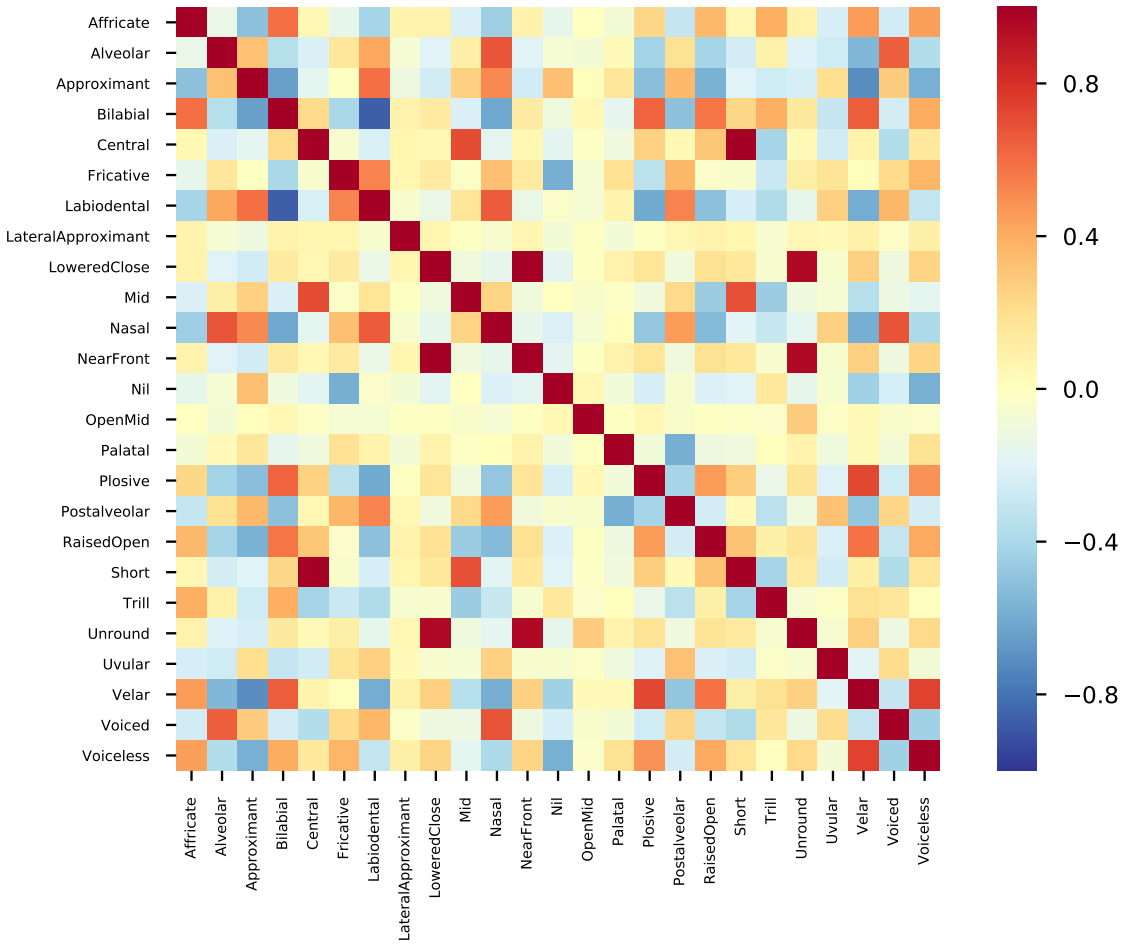



Abbildung 4.30: Korrelationsmatrix der phonetischen Eigenschaften zu den Observationen der westgermanischen Konsonanten.

Die Korrelationsmatrix in Abbildung 4.30 gibt einen ersten Einblick in die gegensätzlichen Lauteigenschaften und verrät auch, welche vokalischen Laute in dem Datenset zum westgermanischen Konsonantismus vorkommen. Die starke Korrelation von *LowerdClose*, *Unround* und *NearFront* lässt auf einen [ɪ]-Laut schließen, *Central*, *Short* und *Mid* auf das [ə]. Das Fehlen der *Front*- und *Back*-Eigenschaften lässt von der Eigenschaft *OpenMid* auf [ɜ] und von *RaisedOpen* auf [ɐ] schließen. Bei den konsonantischen Eigenschaften gibt es eine deutliche Antikorrelation zwischen *Labiodental* und *Bilabial* sowie *Velar* und *Labiodental*. Der *Plosive-Fricative*-Gegensatz, der bezeichnend für die *dat/das*-Grenze ist, ist zwar vorhanden, aber nicht so stark, wie man vermuten könnte.

Die Hauptkomponentenanalyse reduziert die Dimensionen auf 17, wobei die erste Dimension 30% der Varianz erklärt. Die einflussreichsten Features sind *Bilabial*, *Velar* und *Labiodental*. Des Weiteren werden *Approximant*, *Na-*

*sal* und *Plosive* hoch gewichtet. Eine genaue Übersicht liefert Abbildung 4.31.

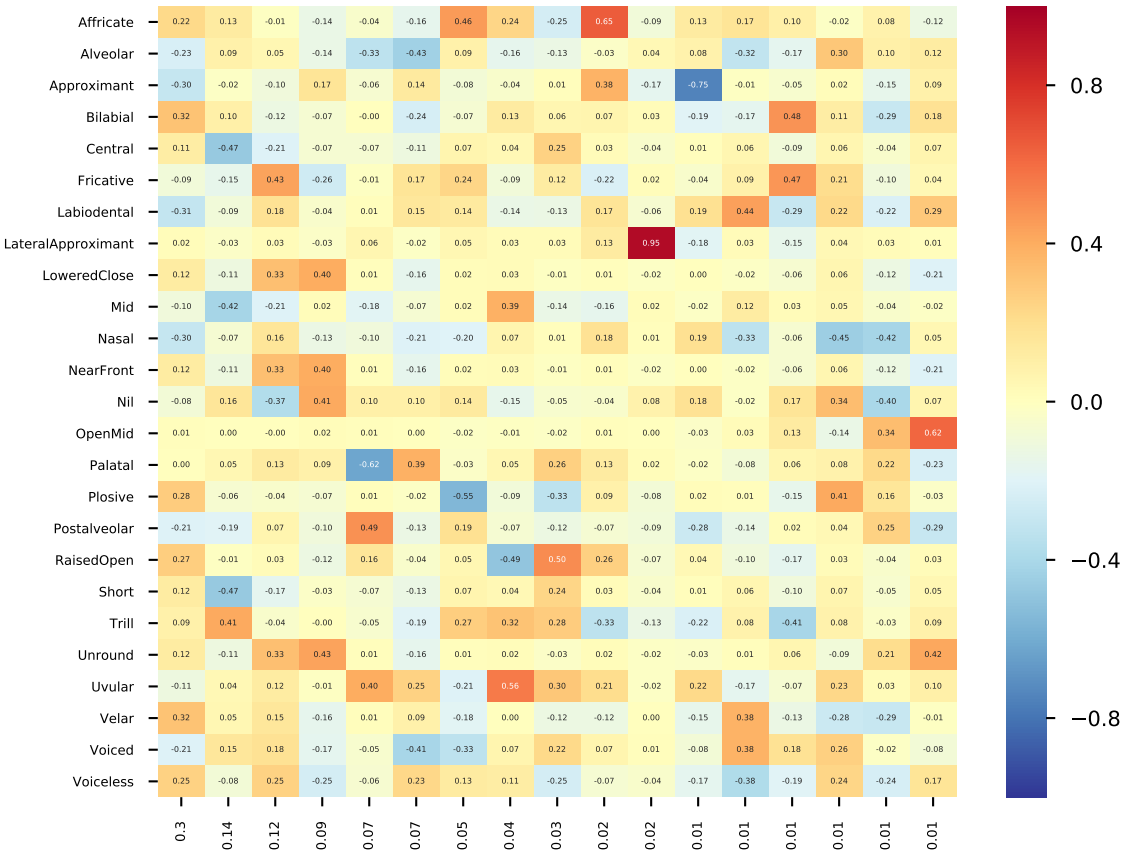



Abbildung 4.31: Anteile der Varianz der ursprünglichen Dimensionen des Konsonantismusdatensets auf die Varianz der neuen, reduzierten Dimensionen nach einer Hauptkomponentenanalyse.

Abbildung 4.32 zeigt die Raumverteilung anhand der PCA. Man sieht eine Violett-Blau-Dominanz im Gebiet des RHEINFRÄNKISCHEN und ein grünbraun markiertes MOSELFRÄNKISCH. Im MOSELFRÄNKISCHEN tritt aber auch ein Gebiet zwischen der *Korf/Korb*- und der *dat/das*-Grenze als etwas stärker türkis hervor. Das UMLAUTGEBIET fällt nicht besonders auf.

*Clusteranalyse*

Ein Zweierclustering (Abbildung 4.33a) zeigt eine interessante Abweichung von dem Gesamtdatenset. Anstelle der *dat/das*-Isoglosse koinzidiert die Clustergrenze mit der *Korf/Korb*-Isoglosse. Der Silhouettenkoeffizient von 0.3, ein Calinski–Harabasz-Wert von 254.99 und ein ARI von 0.99 sprechen für ein stabiles Clustering. Eine Erhöhung auf drei Cluster präsentiert ein ganz anderes Bild (Abbildung 4.33b). Als neues Cluster treten nur ein paar verteilte Orte im RHEINFRÄNKISCHEN hervor, und die Anzahl der Orte mit negativen Silhouetten nimmt im 1-Clustergebiet deutlich zu. Ein Blick in die Daten zeigt, dass der entscheidende Unterschied zwischen dem 1- und dem

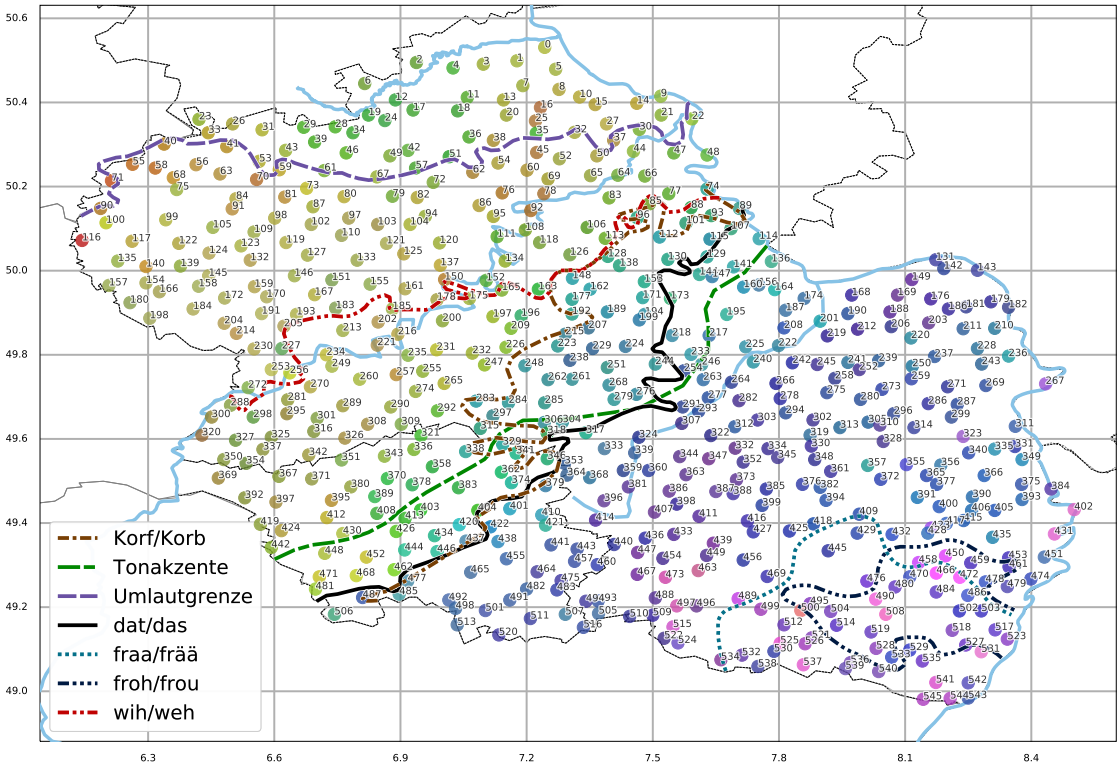
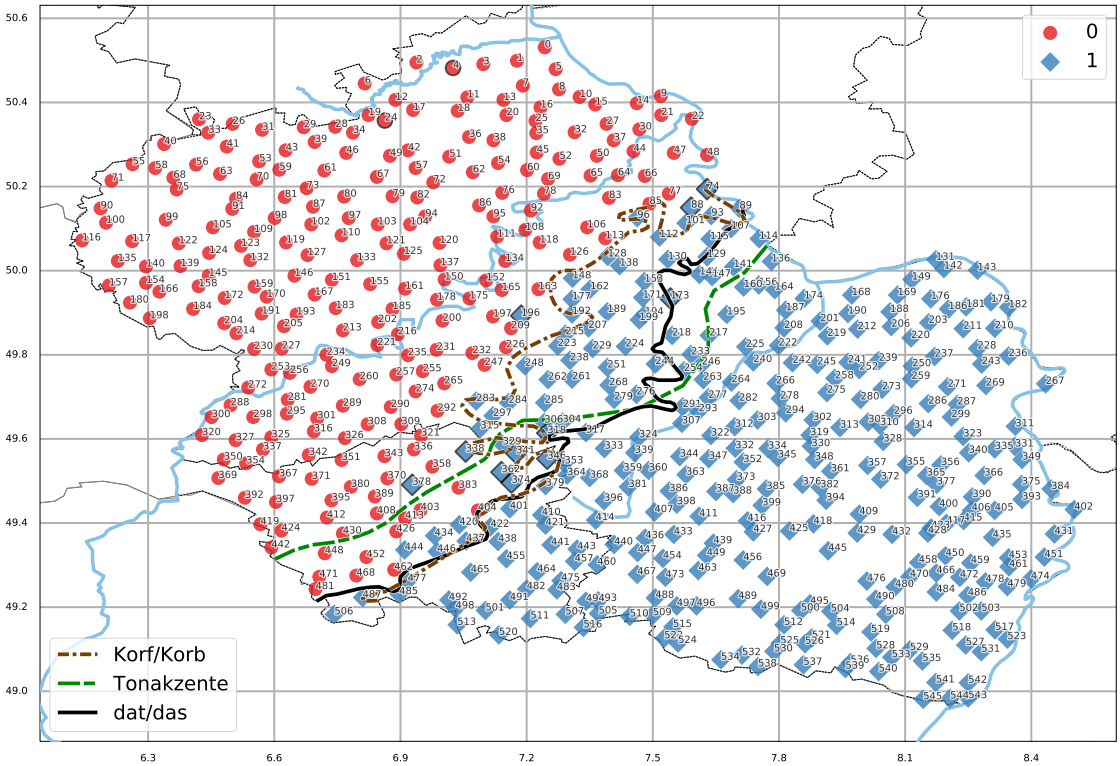



Abbildung 4.32: Räumliche Visualisierung des Datensets zu den westgermanischen Konsonanten durch die ersten drei Dimensionen einer PCA, eingefärbt nach dem HSV Farbmodell.

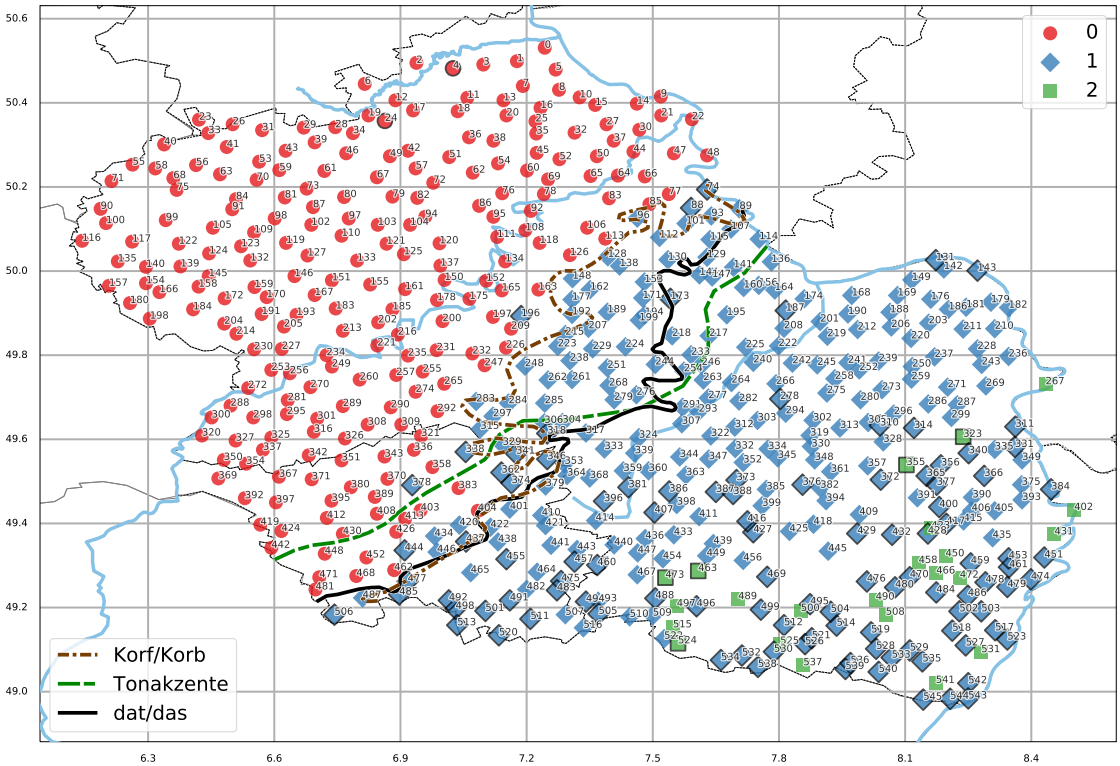
2-Cluster das Auftreten der Eigenschaften des [ɪ]-Lautes in *wg. r* ist. Tatsächlich wird an diesen Orten das Pivotwort *Berg* mit [ɪ] realisiert („Berg“, MRhSA:4/463[160]). Da dieses Phänomen eine sehr starke Abweichung von der Verteilung in dem übrigen Datenset darstellt, führt dies zu einem sehr starken Outliner im Clustering und überschattet damit alle möglichen anderen Strukturen in dem Datenset. Das bedeutet, dass aussagekräftige Clusterings für ein k>2 auf diesem Datenset nicht möglich sind. Für höhere Cluster müssen daher die Daten gefiltert werden, zum Beispiel durch eine Einschränkung auf nur konsonantische Laute oder durch eine Exklusion von *wg. r*.

Der starke Einfluss von [ɪ] in *wg. r* lässt sich auch in Abbildung 4.34 im 2-Cluster sehen. Für die Cluster 0 und 1 ergibt sich, wie zu erwarten, ein komplementäres Bild. So zeigt sich der *Labiodental/Fricative-Bilabial/Plosive*-Gegensatz in *wg. b* als einflussreiches Merkmal für die Trennung von Cluster 0 und 1, und die Hauptgrenze zwischen dem 0- und dem 1-Cluster ähnelt der *Korf/Korb*-Isoglosse. Über alle Daten präsentiert sich der *Fricative* im 0-Cluster als einflussreicher im Vergleich zu *Plosive*. Im 1-Cluster ist es umgekehrt. Die *dat/das*-Isoglosse wird in dieser Clusteranalyse nicht als Hauptgrenze zwischen den zwei größten Clustern erkannt. Genauer gesagt basieren die Cluster sogar auf den gegensätzlichen Features, als es die *dat/das*-

160 <https://www.regionalsprache.de/SprachGis/VectorMap/mrhsa/4/463>, abgerufen 02.02.2018.




(a) KMEANS2




(b) KMEANS3

Abbildung 4.33: KMEANS2- (a) und KMEANS3-Clustering (b) für das Datenset der Lautklassen des westgermanischen Konsonantismus.

Isoglosse implizieren würde. Dass die *Korf/Korb*-Isoglosse dominanter gegenüber der *dat/das*-Isoglosse als Hauptgrenze zwischen den beiden Clustern ist, ist allerdings leicht zu erklären. Da es keine Gewichtung der Features gibt, werden alle Features als gleichwertig betrachtet. Die Laute [f] und [b] unterschieden sich in zwei Features, während [t] und [s] sich nur in der Artikulationsart unterscheiden. Dies führt dazu, dass das Clustering eine bessere Trennung entlang der *Labiodental/Fricative-Bilabial/Plosive*-Hyperebene erzielen kann.

*Höhere Clusterings*

Mit einem gefilterten Datenset (in diesem Fall eingeschränkt auf konsonantische Lauteigenschaften) lassen sich auch höhere Cluster betrachten. Das gefilterte Dreiercluster (Abbildung 4.35a) zeigt eine Abspaltung einer Region im Süden ähnlich dem SÜDPFÄLZISCHEN RELIKGEBIET. Ein Alleinstellungsmerkmal des neuen 2-Clusters ist die hohe Frequenz von *Fricative* in *wg. g*. Der Silhouettenkoeffizient von 0.30 und der Calinski-Harabasz-Wert von 221.77 lassen auf eine ähnliche Stabilität wie KMEANS2 schließen. Das Viercercluster (Abbildung 4.35b) erzeugt ein Zwischengebiet, das sich im Norden zwischen der *Korf/Korb*-Grenze und der *dat/das*-Grenze sowie im Süden nördlich der *dat/das*-Grenze und nördlich der Tonakzentgrenze aufspannt, also entlang der bereits bekannten Hauptgrenze aus dem ALLE-Experiment. Dieses Cluster (1-Cluster) zeichnet sich durch eine hohe Frequenz des *Trill* in *wg. d*, *wg. r*, *wg. s*, *wg. t* und *wg. þ* aus. Dieses Phänomen ist aber sehr wortgebunden und nur die Kombination von Observationen mehrerer Karten lässt ein Cluster entstehen. Mit einem Silhouettenkoeffizienten von 0.26 und einem Calinski-Harabasz-Wert von 189.04 ist dieses Clustering zwar weniger stabil als KMEANS3, aber immer noch stabiler als andere Clusterings. Der ARI ist mit 0.84 ebenfalls hoch. Ein Bootstrapping (siehe Abschnitt A.5, Abbildung A.4b auf Seite 221) rekonstruiert die Ausgangscluster sehr stabil, nur für das 1-Cluster zeigt sich ein gewisser Einfluss des 2-Clusters. Dadurch kann dieses Zwischencluster wieder als Übergangsgebiet gesehen werden, in diesem Fall aber ausgehend von dem 2-Cluster. Ein Fünferclustering (Abschnitt A.5, Abbildung A.4a auf Seite 221) liefert für das Gebiet des MOSELFRÄNKISCHEN kein deutlich zusammenhängendes Clustering mehr. Es spaltet sich allerdings wieder ein Gebiet in der Westeifel ab.

*Merkmaleinfluss*

Die wichtigsten Eigenschaften für die verschiedenen Clusterings sind in Tabelle 4.4 aufgeführt, dabei ist zu beachten, dass nur KMEANS2 auch vokalische Eigenschaften berücksichtigt. Man sieht, dass die Bewertung der Eigenschaften für die verschiedenen Clusterings ähnlich ist. Tabelle 4.4 ist zu entnehmen, dass *Bilabial*, *Labiodental* und *Velar* die wichtigsten Eigenschaften sind. Damit verbunden sind die Laute [b] und [f] der *Korf/Korb*-Grenze. In KMEANS2 sieht man mit *RaisedOpen* zudem noch eine hoch gewichtete vokalische Eigenschaft. Diese Eigenschaft kann in diesem Datenset nur dem [ɐ]-Laut zugeordnet sein, der nur als Realisierung von *wg. r* auftritt.

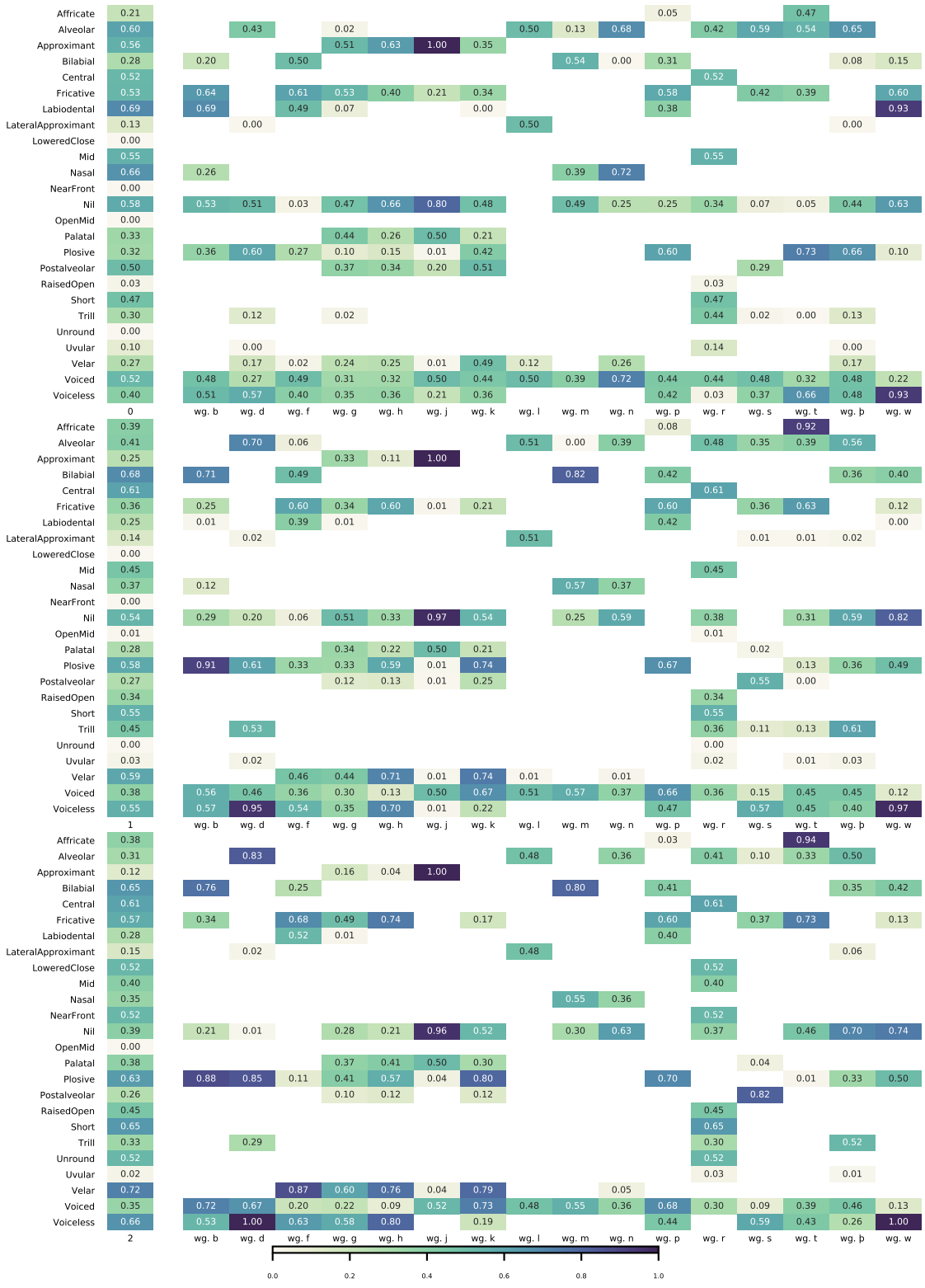
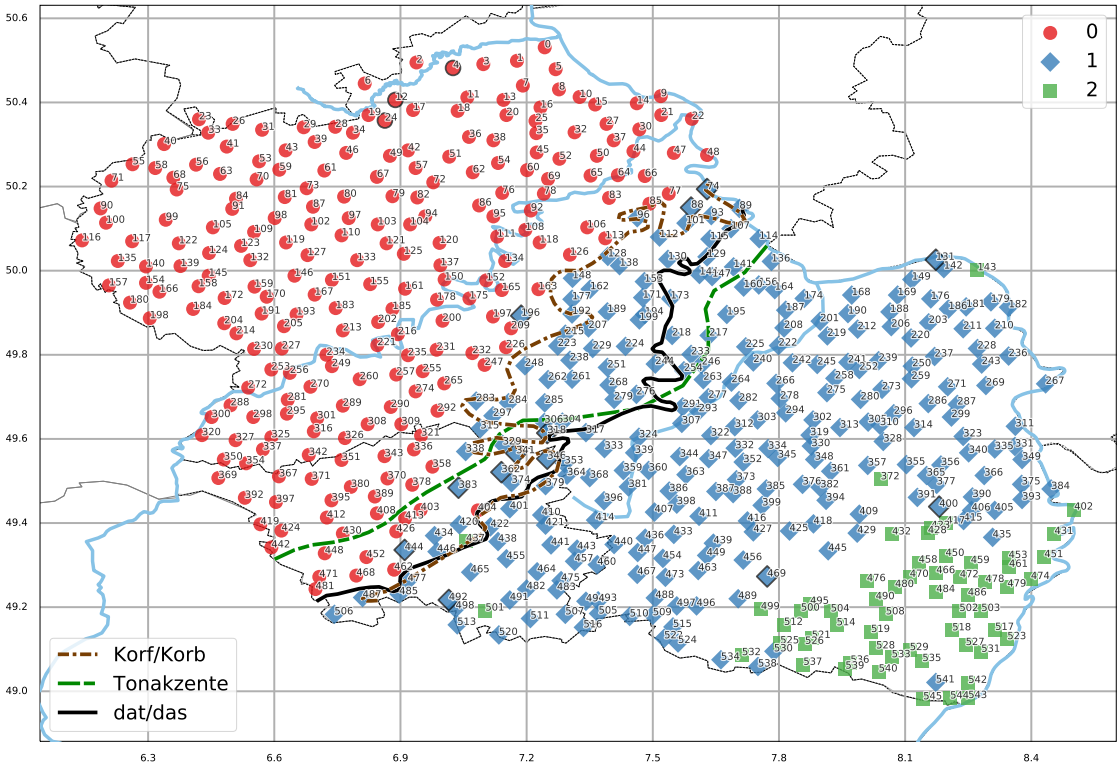
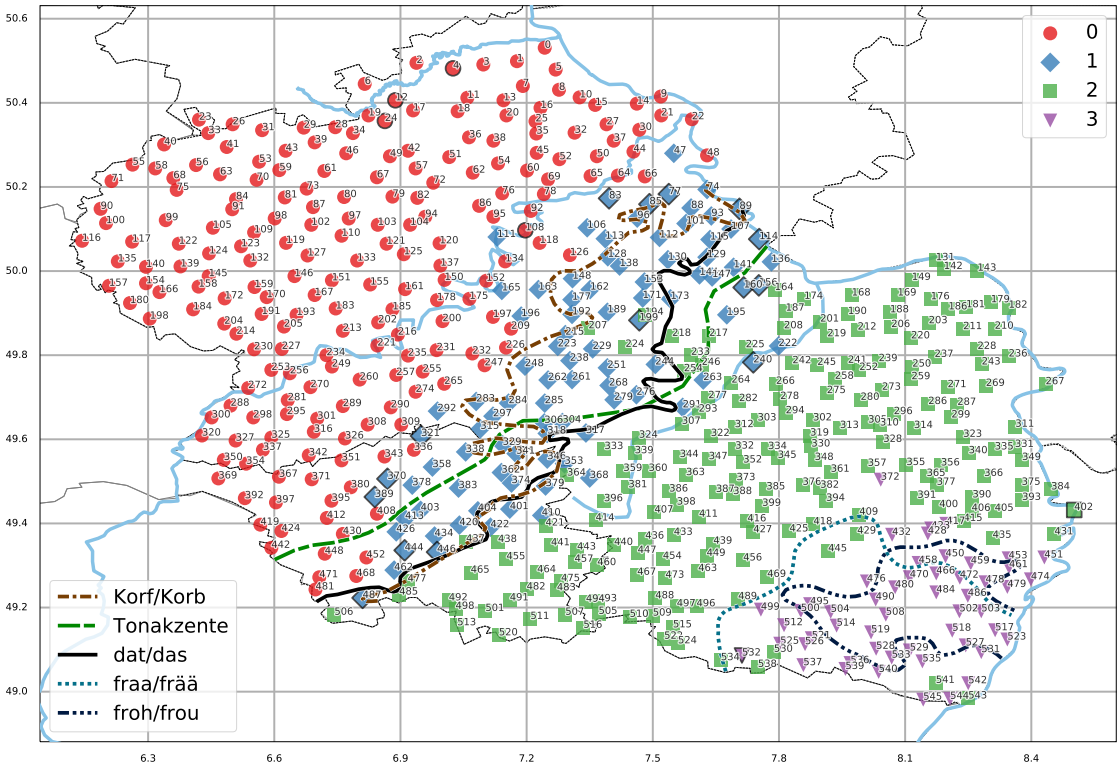



Abbildung 4.34: Mittlere Verteilung der einzelnen phonetischen Eigenschaften nach Cluster und aufgeteilt nach historischen Lautklassen und den westgermanischen Konsonanten für KMEANS3.




(a) KMEANS3




(b) KMEANS4

Abbildung 4.35: KMEANS3- (a) und KMEANS4-Clustering (b) für das Datenset der Lautklassen des Westgermanischen mit Einschränkung auf konsonantische Lauteigenschaften.

Da in den höheren Clustern die vokalischen Eigenschaften herausgefiltert sind, kann keine Aussage mehr zum Einfluss dieser Eigenschaften auf die Sprachraumstruktur bei höheren Clusterings getroffen werden.

Tabelle 4.4: Die zehn höchstsignifikanten ($p$-value < 0.001) Eigenschaften für verschiedene Clusterings auf dem Datenset für die westgermanischen Konsonanten. KMEANS3 und KMEANS4 basieren auf dem auf konsonantische Eigenschaften gefilterten Datenset. Für KMEANS2 wird *OpenMid* und für KMEANS4 wird *LateralApproximant* als nicht signifikant angesehen.

| KMEANS2 | KMEANS3 | KMEANS4 |
|---|---|---|
| Bilabial | Bilabial | Bilabial |
| Labiodental | Labiodental | Velar |
| Velar | Velar | Labiodental |
| Approximant | Approximant | Approximant |
| Nasal | Voiceless | Plosive |
| Plosive | Fricative | Voiceless |
| Affricate | Nasal | Nasal |
| RaisedOpen | Plosive | Trill |
| Voiceless | Affricate | Fricative |
| Alveolar | Nil | Affricate |

Abbildung 4.36 zeigt die Aufteilung nach den westgermanischen Lautklassen für KMEANS4. Das 0-Cluster ist bereits in Abbildung 4.34 gezeigt. Das neu hinzugekommene 1-Cluster zeichnet sich zum einen durch den Wegfall des Konsonanten (siehe *Nil* bei *wg. b, wg. g, wg. j, wg. k, wg. n* und *wg. w*) aus und zum anderen durch eine hohe Frequenz an *Trill* bei *wg. d, wg. r* und *wg. þ*. Im 2-Cluster sind natürlich *Plosive* in *wg. b* und *Fricative* in *wg. t* einflussreich. Die *Nil*-Eigenschaft verhält sich ähnlich wie im 1-Cluster. Das 3-Cluster zeichnet sich durch eine deutlich höhere *Fricative*-Frequenz in *wg. f, wg. g* und *wg. h* aus. Insgesamt nimmt die Frequenz von *Velar* vom 0- bis zum 3-Cluster zu. Dies entspricht einer Zunahme von Nordosten nach Südwesten.

*Bemerkungen*

Anders als es vielleicht von einem konsonantischen Datenset zu erwarten gewesen wäre, trennen sich die Hauptcluster nicht an der *dat/das*-Isoglosse, sondern an der *Korf/Korb*-Isoglosse, was dazu führt, dass die Merkmalshäufigkeiten von *Plosive* und *Fricative* vertauscht sind. Das nördliche Gebiet, welches den größten Bereich des MOSELFRÄNKISCHEN abdeckt, hat mehr *Fricative*-Eigenschaften als das südliche Gebiet. Für *Plosive* ist es genau umgekehrt. Die *dat/das*-Isoglosse wird erst in einem höheren Clustering zu einer Clustergrenze. Dabei muss allerdings beachtet werden, dass das Clustering auf

gleichgewichteten Features basiert und damit nicht einer wahrnehmungsgebundenen Einteilung entsprechen muss. Als eines der häufigsten Worte im Deutschen ist „das“[161] mit einer Häufigkeitsklasse von 2 im Sprachgebrauch natürlich deutlich frequenter als zum Beispiel „Korb“[162] mit einer Häufigkeitsklasse von 12. Ein großes Problem bei dieser Clusteranalyse ist *wg. r*. Diese Lautklasse wird teilweise durch vokalische Eigenschaften realisiert. Diese klare Abweichung von dem ansonsten auf konsonantische Eigenschaften fokussierten Datenset führt zu gravierenden Problemen bei einem Clustering mit k > 2. Dies tritt auf, weil die vokalischen Eigenschaften nur in Observationen zu *wg. r* vorkommen, wohingegen konsonantische Eigenschaften über alle Lautklassen verteilt sind, was dazu führt, dass die vokalischen Laute sehr isoliert sind und folglich als eigenständiges Cluster hervortreten.

Ein Clustering mit k = 4 führt zu einer Raumstruktur, die sich mit dem Moselfränkischen, dem mosel-rheinfränkischen Übergangsgebiet, dem Rheinfränkischen und dem südpfälzischen Reliktgebiet überdecken und bietet damit ein bereits bekanntes Raumbild.

161 <http://corpora.uni-leipzig.de/de/res?corpusId=deu_newscrawl_2011&word=das>, abgerufen 20.05.2018.

162 <http://corpora.uni-leipzig.de/de/res?corpusId=deu_newscrawl_2011&word=Korb>, abgerufen 20.05.2018.

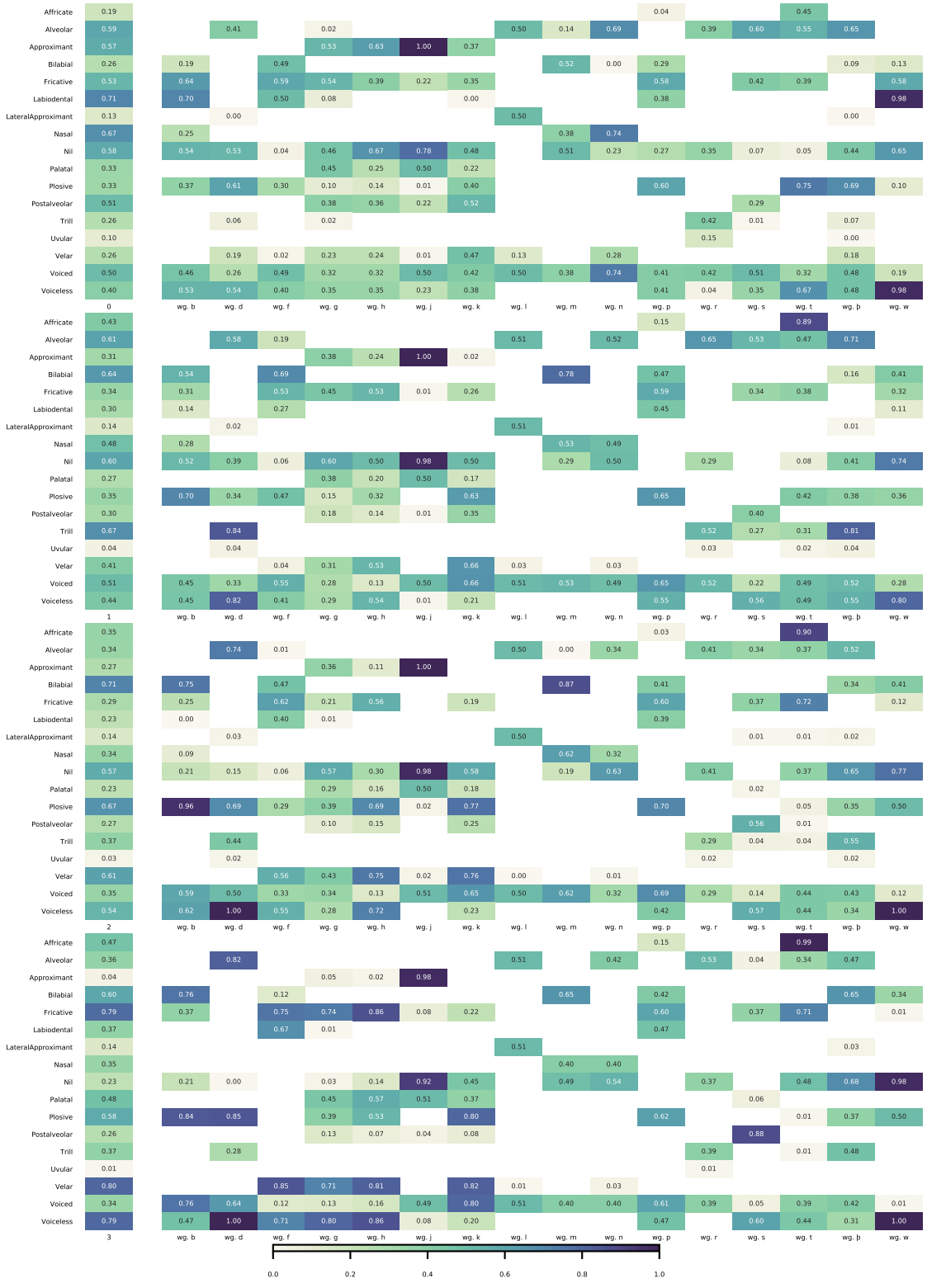



Abbildung 4.36: Mittlere Verteilung der einzelnen phonetischen Eigenschaften nach Cluster und aufgeteilt nach historischen Lautklassen und den westgermanischen Konsonanten für KMEANS4 auf dem auf konsonantische Eigenschaften eingeschränkten Datenset.

## 4.6 DISKUSSION

Die vorgestellten Experimente geben einen Einblick in die Struktur der im Mittelrheinischen Sprachatlas erfassten Daten, die durch die *phonOntology* in ein für eine statistische Datenanalyse geeignetes Format transformiert wurden. Die Daten sind nicht direkte Audiosignale, sondern bereits normalisierte IPA-Notationen, die in ein historisches Referenzsystem gesetzt sind und in Form von Sprachkarten publiziert wurden. Die *phonOntology* erzeugt aus diesen Notationen Lauteigenschaften. Aus diesen Lauteigenschaften und ausgewählten Teilmengen des historischen Bezugssystems werden die Datensets für die Experimente generiert. Diese Teilmengen folgen gängigen linguistischen Betrachtungen und lassen sich unterteilen in das gesamte Bezugssystem mit allen Referenzlauten, die Bezugslaute, die nur dem mittelhochdeutschen Langvokalismus zugeordnet sind, die Bezugslaute, die nur dem mittelhochdeutschen Kurzvokalismus zugeordnet sind und alle Bezugslaute zum westgermanischen Konsonantismus. Nicht näher aufgeführt sind Experimente zu den gesamten mittelhochdeutschen Vokalen und eingeschränkte Experimente auf die mittelhochdeutschen Diphthonge. Die so generierten Datensets werden mittels einer Clusteranalyse strukturiert. Dabei kommen drei Clusteringmethoden zum Einsatz, die je zwei bis fünf Cluster erzeugen. Alle drei Clustermethoden haben gewisse Stärken und Schwächen und die Wahl des Clusteralgorithmus kann deutlichen Einfluss auf die Form der Cluster haben. Das Gaussian Mixture Model versucht in den Daten k Normalverteilungen zu finden, dieses Verfahren basiert auf Wahrscheinlichkeiten und eignet sich besonders für Daten, die auch einer Normalverteilung folgen. Da es aber auf der Einbettung von Gaußs verteilungen basiert, hat der Algorithmus Schwierigkeiten, kompakte Daten oder Daten, die nur über schwache Cluster verfügen, klar zu trennen. Der K-Means-Algorithmus erzeugt Clusterzentren und ordnet die Datenpunkte den nächsten Zentren zu. Der Algorithmus funktioniert gut bei konvexen Datensets, aber er hat Probleme mit Ausreißern und komplexeren Datenstrukturen. Das Ward-Clustering erzeugt Cluster mit ähnlicher Varianz. Dies erlaubt auch eine Anwendung auf komplexere Datenstrukturen. Allerdings tendiert dieser Algorithmus dazu, ausgeglichene Cluster zu erzeugen, was nicht unbedingt den darunter liegenden Daten entsprechen muss. Da eine Clusteranalyse ein unüberwachter Lernalgorithmus ist, steht kein explizites Referenzdatenset zur Verfügung, gegen das das geclusterte Datenset getestet werden kann. Dies macht es schwer, von richtigen oder falschen Clusterings zu sprechen. Stattdessen wird die Güte eines Clusterings an der internen und Zwischenclusterstreuung bewertet (Silhouettenkoeffizient und Calinski-Harabasz-Wert) sowie der Stabilität (Adjusted-Rand-Index) gegenüber zufälligem Aussortieren von einem Teil der Daten und Bootstrapping. Ein weiteres Qualitätsmerkmal, das allerdings bereits auf einer subjektiven Interpretation beruht, ist der räumliche Zusammenhang der den Clustern zugeordneten Ortspunkte. Dies beruht auf der Annahme, dass eine Kausalität zwischen der Ähnlichkeit der Datenpunkte und der räumlichen Nähe der zugehörigen Ortspunkte besteht. Die in diesem Kapitel als Sprachraumeinteilung ausgewählten Sprachräume basieren auf den Clusterings, die unter den drei vorgestellten Bewertungsgesichtspunkten am besten abschneiden. In vielen Fällen sind die Unterschiede

zwischen verschiedenen Clusteralgorithmen aber gering. Tabelle 4.5[163] gibt eine Übersicht über den V-Measure-Wert[164] (vgl. Rosenberg und Hirschberg 2007) zwischen den verschiedenen Clusteralgorithmen. Man sieht eine hohe Übereinstimmung zwischen GMM und WARD mit einem K von 3 und die niedrigsten Werte bei KMEANS und auch bei einem k von 3 für das ALLE-Experiment. Während GMM und WARD Cluster erzeugen, die einer räumlichen Einteilung in das nördliche UMLAUTGEBIET (0-Cluster), das MOSELFRÄNKISCHE (1-Cluster) und das RHEINFRÄNKISCHE (2-Cluster) ähneln, findet KMEANS eine andere Einteilung für das 0- und 1-Cluster. In diesem Fall wird ein Teil der Westeifel noch mit zu dem 0-Cluster gezählt. Dadurch entsteht aber ein Cluster negativer Silhouettenkoeffizienten an vielen Orten.

Tabelle 4.5: V-Measure-Wert zwischen den Clusteralgorithmen für die einzelnen Experimente.

| | VERGLEICH | K = 2 | K = 3 | K = 4 | K = 5 |
|---|---|---|---|---|---|
| ALLE | GMM vs. KMEANS | 0.89 | 0.69 | 0.61 | 0.70 |
| | KMEANS vs. WARD | 0.87 | 0.68 | 0.60 | 0.79 |
| | WARD vs. GMM | 0.87 | 0.90 | 0.83 | 0.71 |
| LANG | GMM vs. KMEANS | 0.42 | 0.40 | 0.78 | 0.60 |
| | KMEANS vs. WARD | 0.86 | 0.62 | 0.69 | 0.69 |
| | WARD vs. GMM | 0.40 | 0.54 | 0.64 | 0.69 |
| KURZ | GMM vs. KMEANS | 0.67 | 0.67 | 0.51 | 0.54 |
| | KMEANS vs. WARD | 0.73 | 0.70 | 0.64 | 0.73 |
| | WARD vs. GMM | 0.84 | 0.85 | 0.68 | 0.69 |
| WG | GMM vs. KMEANS | 0.97 | 0.72 | 0.64 | 0.64 |
| | KMEANS vs. WARD | 0.69 | 0.58 | 0.76 | 0.66 |
| | WARD vs. GMM | 0.68 | 0.58 | 0.59 | 0.61 |

Ein Beispiel dafür, dass ein Clusteralgorithmus nicht in der Lage ist, räumliche Cluster zu erzeugen, ist der GMM2, angewendet auf das Datenset zu den westgermanischen Konsonanten (WG). Wie in Abschnitt 4.5 erwähnt, führen die vokalischen Eigenschaften bei *wg. r* zu Problemen beim Clustering. Beim GMM2-Algorithmus führt dies dazu, dass Orte, bei denen das /r/ in „Berg“ mit einem /ɪ/ realisiert wird, ein Cluster bilden und die restlichen Orte das andere Cluster. Solche Ergebnisse sind zwar gut geeignet, um auf Probleme oder Besonderheiten in einem Datenset aufmerksam zu machen,

163 Für WG wird das Datenset verwendet, welches auf die konsonantischen Eigenschaften eingeschränkt ist. Das Datenset, welches alle Laute zulässt, ist für GMM2 nicht in der Lage, eine Trennung der Daten vorzunehmen. Dies resultiert zum Beispiel in einem V-Score von 0.06 für GMM2 vs. KMEANS2.

164 Der V-Measure-Wert ist ein entropiebasiertes Ähnlichkeitsmaß, welches die Überdeckung von Elementen zwischen zwei Mengen misst. Eine spezifische Labelzuordnung spielt dabei keine Rolle. Ein Wert von 1 bedeutet eine perfekte Überdeckung.

als Sprachraumeinteilung sind die dermaßen erzeugten Cluster allerdings ungeeignet. Niedrige V-Measure-Werte (<0.6) sprechen für eine unterschiedliche Raumeinteilung. Man sieht auch, dass sich die Werte für k=5 über alle Experimente hinweg etwas normalisieren. Clustering kann somit auch als ein Entdeckungsvorgang über die Daten verwendet werden. Bei den Experimenten haben sich immer wieder ähnliche Regionen herausgebildet, teilweise aber in einer anderen Reihenfolge bei Erhöhung des Parameters k. So zeigt sich immer wieder ein Gebiet in der Westeifel (vgl. Abbildung 4.17b auf Seite 108 und Abbildung 4.26 auf Seite 121), das bei verschiedenen k, Clusteralgorithmen und Datensets auftaucht. Auch treten Bereiche im SÜDPFÄLZISCHEN RELIKTGEBIET bei verschiedenen Experimenten hervor (vgl. Abbildung 4.9 auf Seite 97, Abbildung 4.26 auf Seite 121 und Abbildung 4.35b auf Seite 132). Solche Zusammenhänge können als Ausgangspunkt genauerer Untersuchungen dieser Teilregionen gesehen werden. Im Fall der Westeifel lässt sich zum Beispiel eine Vertauschung der Eigenschaft *Long* und *Short* zwischen dem Langvokal- und dem Kurzvokaldatenset zeigen. Bei den besprochenen Experimenten wurden Clusteranalysen für $k \leq 5$ aufgeführt, da ab einer gewissen Clusteranzahl die Idee des Sprachraums nicht mehr gegeben ist und sich Cluster basierend auf leichten Abweichungen oder nur eines einzelnen Merkmals herausbilden. Auch sinkt bei hohem k die Stabilität der Cluster bei einer Auswahl zufälliger Teilmengen ab und macht damit eine Bewertung der Cluster schwieriger bzw. weniger aussagekräftig.

## 4.7 ZUSAMMENFASSENDE BEOBACHTUNGEN

Die Experimente offenbaren räumliche Strukturen, die in vielen Fällen mit den historischen Sprachräumen übereinstimmen. Damit können die entsprechenden Cluster als Repräsentanten für den entsprechenden Sprachraum dienen und die darunter liegenden Daten als ein Modell dieses Raums. Ein solches Modell basiert auf den phonetischen Eigenschaften, die in der *phonOntology* definiert wurden. Dadurch erhält man zum einen eine mathematisch nachvollziehbare Basis für die Sprachräume und zum anderen erlaubt es, Datensets, die mittels der *phonOntology* annotiert wurden, untereinander zu vergleichen.

Die durch die Clusteranalyse ermittelten Grenzen zwischen den einzelnen Sprachräumen folgen weitestgehend den in Abschnitt 2.5 besprochenen Isoglossen des *Mittelrheinischen Sprachaltlas*. Allerdings tun sich besonders im MOSELFRÄNKISCHEN ein paar Eigenheiten auf.

### *Das MOSELFRÄNKISCH-RHEINFRÄNKISCHE Grenzgebiet*

Als Hauptgrenzen zwischen dem MOSELFRÄNKISCHEN und dem RHEINFRÄNKISCHEN werden zum einen die *dat/das*-Grenze und zum anderen die Tonakzentgrenze gesehen. Diese doppelte Grenze bestätigt sich in den Experimenten.

Das ALLE-Experiment trennt die Hauptcluster zum MOSELFRÄNKISCHEN und RHEINFRÄNKISCHEN entlang der Tonakzentgrenze und der *dat/das*-Isoglosse (vgl. Abbildung 4.4b auf Seite 89). Diese Grenze findet sich auch bei

den anderen Experimenten, die als Teilmengen des ALLE-Experiments gesehen werden können. Während die Grenze zu den historischen Kurzvokalen des Mittelhochdeutschen sich deutlich stärker an den Tonakzenten orientiert, verschiebt sich die Grenze nach einer Filterung der Tonakzente in Richtung der Hauptgrenze aus ALLE. In den Clusterings zu den Langvokalen und den Konsonanten findet sich diese Grenze ebenfalls, allerdings erst ab einem k > 2. Der Bereich zwischen diesen beiden Isoglossen kann als die Grenz- oder Übergangsregion zwischen dem MOSELFRÄNKISCHEN und dem RHEINFRÄNKISCHEN betrachtet werden. Bootstrapping (siehe Abschnitt A.5, Abbildung A.1b auf Seite 218) zeigt, dass diese Behauptung für den östlichen Bereich deutlicher zutrifft als für den westlichen Bereich. Clusterings mit höherem k zeigen zudem für alle Experimente, dass diese Hauptgrenze stabil bleibt und sich neue Cluster eher im MOSELFRÄNKISCHEN bilden. Das lässt das RHEINFRÄNKISCHE homogener erscheinen. Nur im Süden bildet sich ein Cluster im Gebiet des SÜDPFÄLZISCHEN RELIKTGEBIETES heraus. Im Fünferclustering (siehe Abbildung 4.9 auf Seite 97) zerfällt der Bereich des MOSELFRÄNKISCHEN in ein Cluster zum nördlichen UMLAUTGEBIET, ein mittleres Hauptgebiet sowie ein Zwischengebiet entlang der *Korf/Korb*-Isoglosse und der Hauptgrenze.

Das Bootstrapping zeigt aber, dass dieses Zwischengebiet eher als ein Übergangsgebiet von dem Hauptgebiet aus gesehen werden sollte. Das dem Zwischengebiet zugeordnete 2-Label findet sich mit sichtbaren Anteilen auch an Orten, die im Clustering dem 1-Label zugeordnet sind[165]. Das 1-Label hingegen hat nur Anteile an Orten, die auch dem 1-Cluster zugeordnet sind. Zudem kann man besonders in dem bereits erwähnten Grenzgebiet auch einen geringen Einfluss des 3-Labels auf das Zwischengebiet ausmachen. Auffällig ist, dass es so gut wie keinen Einfluss von 1- oder 2-Label auf die Cluster im Gebiet des RHEINFRÄNKISCHEN gibt.

*Das UMLAUTGEBIET*

In der Clusteranalyse tritt das nördliche UMLAUTGEBIET sehr deutlich hervor. Die dominierende Eigenschaft ist *Round*. Dieses Gebiet erscheint als eigenständiges Cluster ab einem k > 2 bei Clusterings über alle Lauteigenschaften und über die Lauteigenschaften zu den historischen Kurzvokalen des Mittelhochdeutschen. In der in Abbildung 4.37 dargestellten Visualisierung der Daten mittels der T-Distributed-Stochastic-Neighbor-Einbettung entspricht das 0-Cluster dem UMLAUTGEBIET, und man beobachtet in dieser Form der Visualisierung einen Abstand zu den anderen Clustern[166]. In den Clusterings zu den historischen Langvokalen des Mittelhochdeutschen zeigt sich das UMLAUTGEBIET nicht als eigenständiges Cluster, obwohl die *Round*-Eigenschaft auch dort dominant ist. Grund dafür ist der deutlich stärkere Fokus auf die Reihenvertauschung von /e/ - /o/ zu /ɪ, ʏ/ - /ʊ/ in derselben Region.

165 Es ist zu beachten, dass aufgrund der Unschärfe bei bei der Clustermarkierung während des Bootstrappings (siehe Seite 79) der Anteil des 2-Label wahrscheinlich etwas höher ist, als es bei einem streng deterministischen Clustering der Fall wäre.

166 Dabei ist zu beachten, dass TSNE dazu neigt, deutlichere Cluster zu produzieren, und diese Visualisierung nur eine grobe Annäherung an die Struktur der originalen Daten ist.

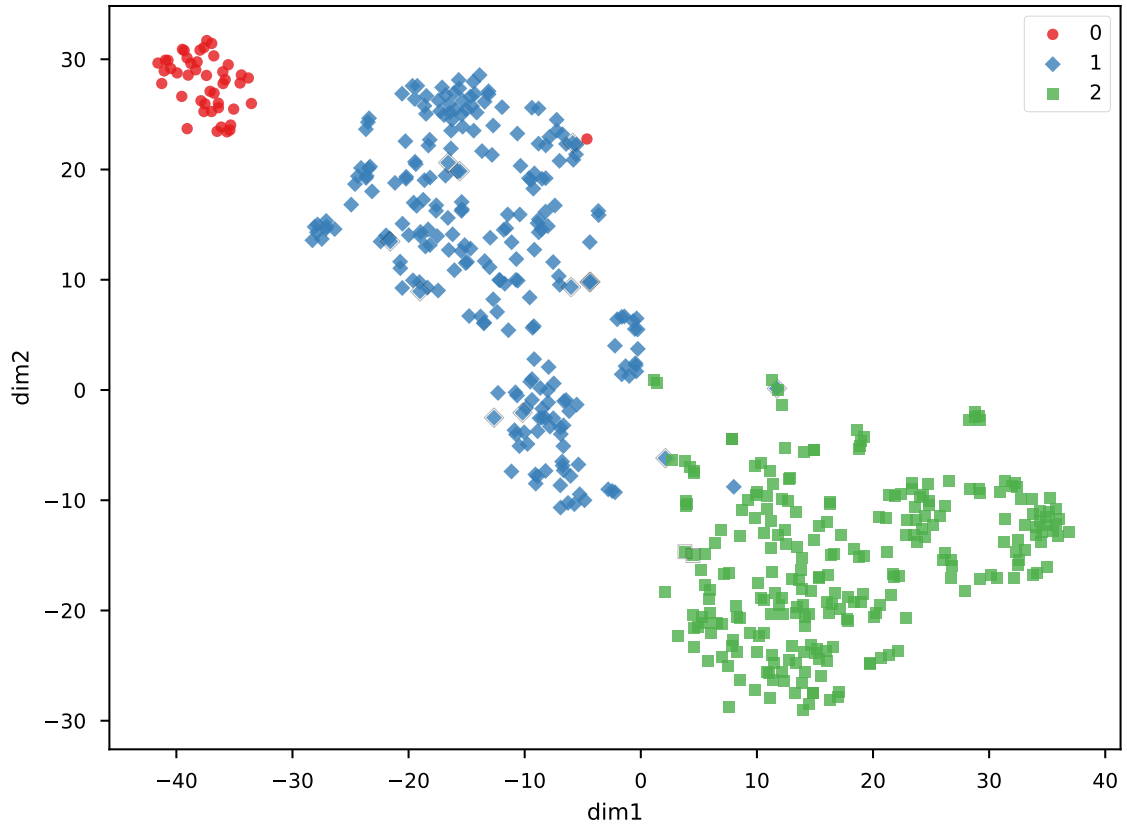



Abbildung 4.37: Visualisierung des ALLE-Datensets mittels einer TSNE und Einfärbung der Datenpunkte nach GMM3.

*Die Grenze der Reihenvertauschung*

Innerhalb des MOSELFRÄNKISCHEN kommt es zu einer Reihenvertauschung von /e/ - /o/ zu /ɪ/ - /ʊ/ bei den historischen Langvokalen des Mittelhochdeutschen. Die Grenze dieser Vertauschung koinzidiert weitestgehend mit der *wih/weh*-Grenze der entsprechenden Wenkerkarte „weh“. Dabei ist im nördlichen Bereich die *LoweredClose*-Eigenschaft und im südlichen die *CloseMid*-Eigenschaft dominant. Diese Reihenvertauschung ist konsequent für dieses Cluster und führt zu einer deutlichen Abgrenzung zwischen den beiden Hauptregionen. Am westlichen Rand des Untersuchungsgebiets findet sich noch eine Häufung von Orten, die dem nördlichen Cluster zugeordnet sind und etwas über die *wih/weh*-Isoglosse hinausgehen, da diese Isoglosse auf dem /ɪ/-/e/-Gegensatz in *mhd. ê* basiert, das Cluster aber auch den /ʊ/-/o/-Gegensatz in *mhd. ô* mit berücksichtigt. Diese Grenze tritt nur in dem Langvokaldatenset auf.

Die Eigenschaften zu den historischen Kurzvokalen verteilen sich über ein deutlich größeres Spektrum, so dass nicht eindeutig von einer Reihenvertauschung gesprochen werden kann, sondern anzunehmen ist, dass eine Realisierung stark von dem gewählten Pivotwort abhängt. Abbildung 4.38 zeigt die Verteilung der Kurzvokale getrennt nach dem Zweierclustering zu

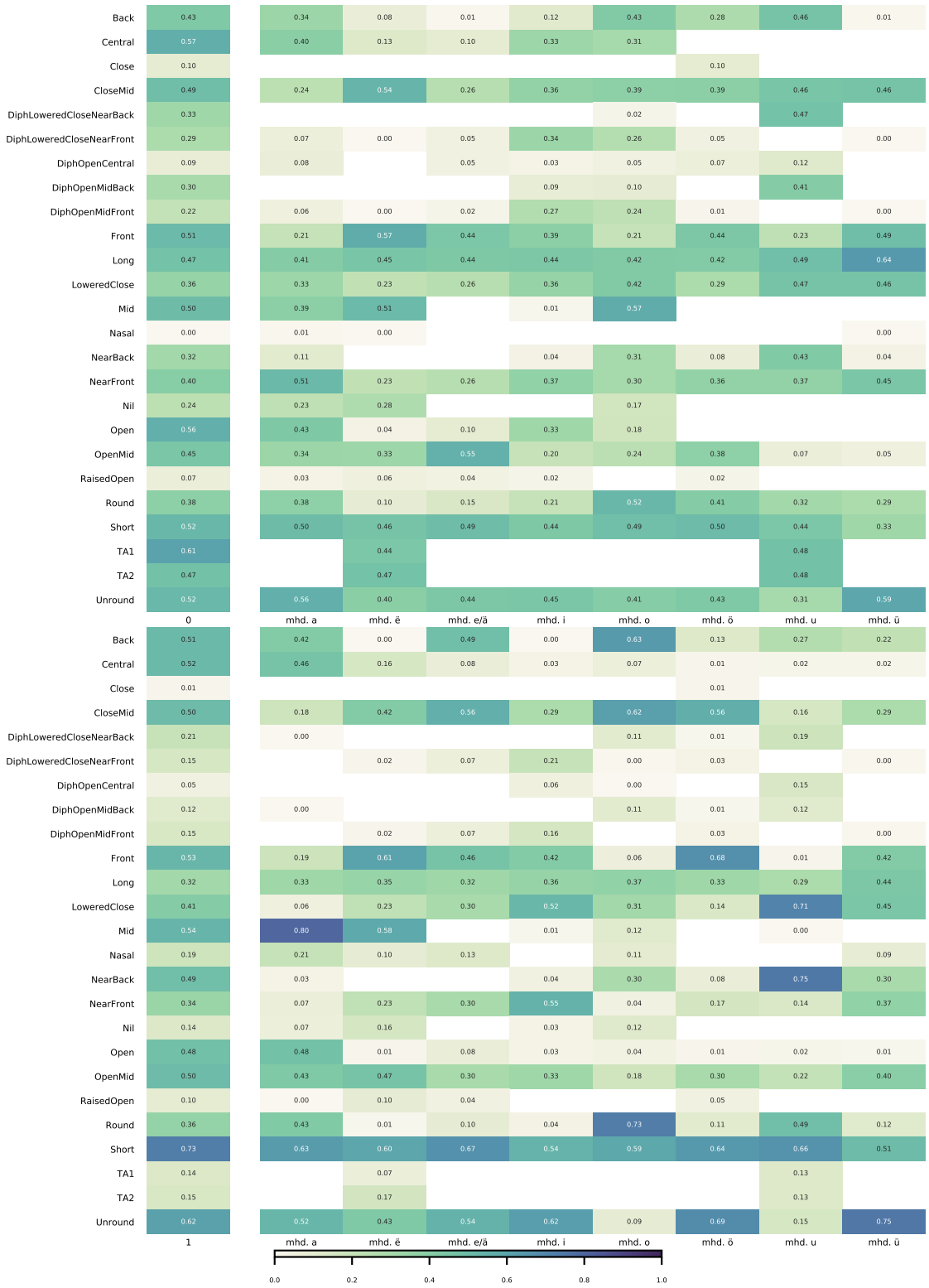



Abbildung 4.38: Verteilung der Lautklassen der historischen Kurzvokale, getrennt nach dem Zweierclustering (KMEANS2) der historischen Langvokale.

den historischen Langvokalen (siehe Seite 103). Man erkennt, dass besonders die Lautklassen *mhd. a, mhd. i, mhd. o, mhd. ö* und *mhd. u* nördlich der *wih/weh*-Isoglosse viel variantenreicher sind als die südlichen Gegenstücke. Eine konsequente Reihenvertauschung ist allerdings nicht zu beobachten.

*Die konsonantische Grenze*

Mit der bereits gezeigten Bedeutung und historischen Signifikanz der *dat/das*-Isoglosse könnte man annehmen, dass diese Grenze auch bei dem Clustering der konsonantischen Lautklassen die Hauptgrenze bildet. Stattdessen teilen sich die Hauptcluster entlang einer Grenze, die der *Korf/Korb*-Isoglosse der Wenkerkarte zu „Korb“ ähnelt. Erst bei einem k > 3 bildet sich eine Grenze entlang der *dat/das*-Isoglosse. Die Clusteranalyse bewertet den *Bilabial-Labiodental*-Gegensatz in *wg. b* höher als den *Fricative-Plosive*-Gegensatz in *wg. t.* Abbildung 4.39 zeigt die mittlere Verteilung der Eigenschaften zu den westgermanischen Konsonanten in den beiden Hauptclustern.

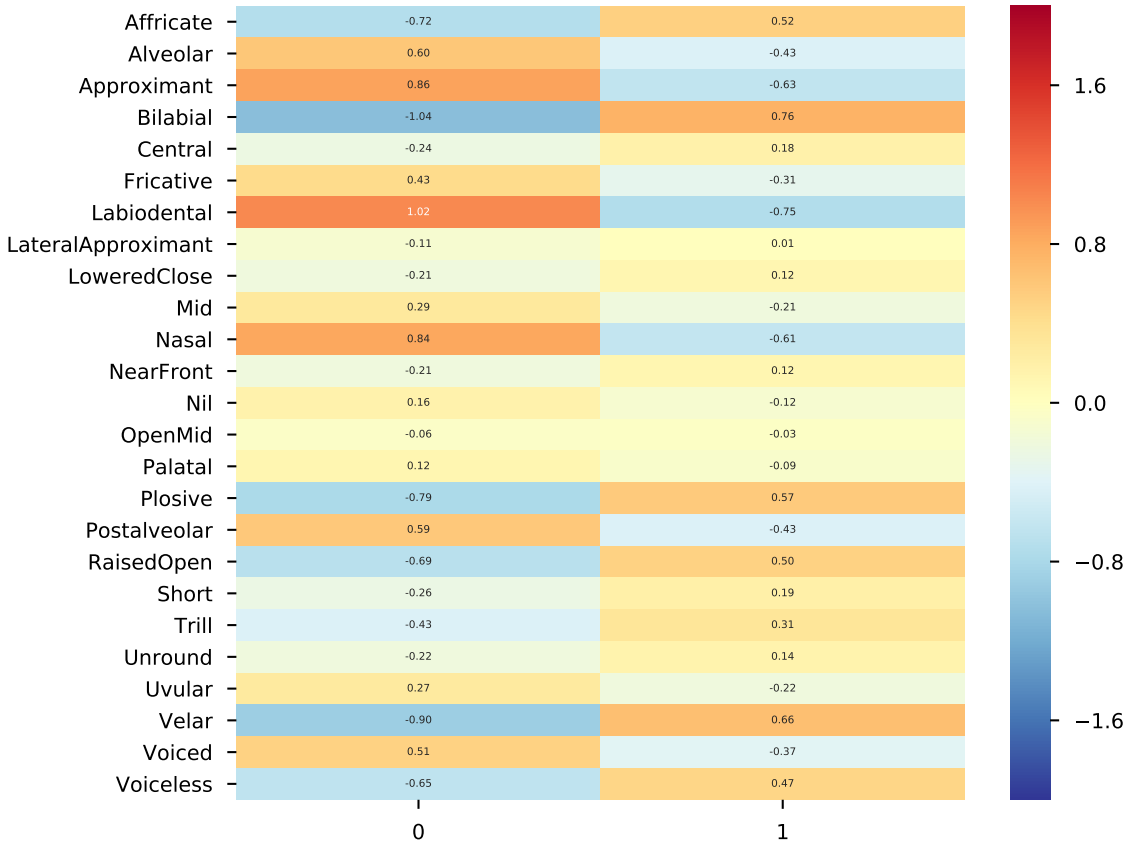



Abbildung 4.39: Mittlere Verteilung der phonetischen Eigenschaften im Datenset zu den westgermanischen Konsonanten nach KMEANS2-Clustering.

Man sieht, dass *Fricative* im 0-Cluster, also im nördlichen Bereich, einen positiven Wert aufweisen und *Plosive* einen negativen. Im 1-Cluster ist die Verteilung vertauscht. Das bedeutet, dass das Vorkommen von *Plosive* im

nördlichen Bereich des Untersuchungsgebietes statistisch seltener ist als im südlichen[167] und *Plosive* in *wg. t* eine Ausnahme im MOSELFRÄNKISCHEN darstellt. Da das Clustering die Grenze entlang der *Korf/Korb*-Isoglosse zieht, ist zu erwarten, dass der nördliche Bereich einen Bias in Richtung *Fricative* hat. Deutlicher werden die Unterschiede aber an einem *Labiodental–Bilabial*-Gegensatz. Das Gebiet zwischen der *Korf/Korb*- und der *dat/das*-Isoglosse entspricht ungefähr dem Zwischengebiet aus dem ALLE-Experiment (mit WARD5). Dieses Zwischengebiet hat kein eindeutiges Erkennungsmerkmal, sondern zeichnet sich eher durch seine hohe Varianz aus und kann als Berührungsraum zwischen dem MOSELFRÄNKISCHEN und RHEINFRÄNKISCHEN interpretiert werden.

*Das SÜDPFÄLZISCHE RELIKTGEBIET*

Das SÜDPFÄLZISCHE RELIKTGEBIET tritt in dem ALLE-Experiment und interessanterweise auch in den untersuchten Teilexperimenten auf, meistens ab einem k von 3 oder 5. Dies ist insofern beachtenswert, als die Hervorhebung dieser Region auf den Karten des MRhSA zu „Frau“ (mhd. ou) und „froh“ (mhd. ô) basiert, die Cluster sich aber auch in den Experimenten zu den historischen Kurzvokalen und den Konsonanten wiederfinden. Bei den konsonantischen Lauten zeichnet sich dieses Gebiet durch ein erhöhtes Vorkommen von *Affricate*, *Fricative* und *Voiceless* im Verhältnis zu dem übrigen Gebiet des RHEINFRÄNKISCHEN aus und bei den Eigenschaften zu den historischen Kurzvokalen zeichnet es sich durch ein erhöhtes Vorkommen an Diphthongeigenschaften und ein etwas niedriges Vorkommen von *Unround* im Verhältnis zu der Umgebung aus. Diese Eigenschaften zu den Kurzvokalen ähneln damit dem Cluster zum UMLAUTGEBIET im Norden des MOSELFRÄNKISCHEN. Diese Nähe zeigt sich auch in den Clusterings wie zum Beispiel in dem KMEANS3-Clustering (Abbildung 4.24b) auf Seite 117. Die Region wird allerdings erst ab einem k = 5 als eigenes Gebiet aufgefasst, davor wird es zu dem o-Cluster gerechnet, welches weitestgehend mit dem UMLAUTGEBIET koinzidiert. Das bedeutet nicht, dass ein direkter Zusammenhang zwischen diesen beiden Räumen besteht. Da die „Nähe“ nur auf sehr wenigen Eigenschaften beruht, kann angenommen werden, dass das Zusammenfassen zu einem Cluster auf einer technischen Limitation des Clusterings beruht. Da zum Beispiel für ein k = 3 exakt drei Cluster gefunden werden, muss dieses Gebiet einem Cluster zugeordnet werden, in diesem Fall dem o-Cluster, da dort die Ähnlichkeit am größten ist.

167 Umgekehrt für *Fricative*, wenn auch nicht so deutlich wie bei *Plosive*.

# 5 VERGLEICH MIT DER JÜNGEREN GENERATION

Als mehrdimensionaler Sprachatlas bietet der *Mittelrheinische Sprachatlas* auch Daten für eine jüngere Generation, also Informanten, die zur Zeit der Erhebung des Atlas ca. 35 Jahre alt waren. Dadurch ist ein zweites Datenset verfügbar, welches Observationen zu dieser Generation enthält und ebenfalls durch die *phonOntology* annotiert werden kann. So können in Form einer „apparent-time"-Analyse Vergleiche zwischen den Generationen durchgeführt werden, die einen Einblick in Änderungen des Sprachraums zwischen diesen beiden Generationen gewähren. Dazu bieten sich verschiedene Methoden an. Eine naheliegende Möglichkeit ist das Durchführen der Clusteranalyse auf dem Datenset zur jüngeren Generation und ein anschließender Vergleich der Cluster. In weiteren Analysen kann das Datenset der älteren Generation als Ausgangsmodell dienen und Klassifikatoren trainieren, welche die jüngere Generation auf Basis der älteren klassifizieren. Auch lässt sich die Distanz zwischen den entsprechenden Datenpunkten beider Datensets berechnen.

Eine Clusteranalyse erzeugt neue Cluster, die in ihrer Bewertung unabhängig von den Clustern basierend auf der älteren Generation sind. Ein direkter Vergleich der Cluster der jüngeren und der älteren Generation ist nur eingeschränkt möglich, da die Bedingungen, die zur Clusterbildung führen, andere sein können. Wenn man aber davon ausgeht, dass die Haupträume weiterhin Bestand haben, sollte ein Vergleich[168] der Hauptcluster durchaus vertretbar sein. Bei der Klassifizierung wird von den Clustern der älteren Generation ausgegangen und Änderungen bei der jüngeren Generation können im Verhältnis zur älteren Generation betrachtet werden. Man ist dabei aber an die Strukturen der älteren Generation gebunden.

## 5.1 VORVERARBEITUNG

Um aussagekräftige Klassifikationen oder Distanzmaße zu erhalten, ist es wichtig bei der Vorverarbeitung der Datensets sicherzustellen, dass beide Datensets aus einer gemeinsamen (hypothetischen) Verteilung kommen. Falls möglich, wird die Verteilung des Trainingsdatensets[169] als Standardverteilung gesetzt und das Testdatenset wird anhand dieser skaliert. Um eine kompatible Skalierung zu gewährleisten, wird zudem das Testdatenset so angepasst, dass es nur auf Karten basiert, zu denen es auch entsprechende Karten für die jüngere Generation gibt. Dies bedeutet zum Beispiel, dass die Tonakzente als Eigenschaften fehlen. Die Zuordnung der Label basiert aber weiterhin auf den durch die Clusteranalyse bestimmten Klassen.

168 Auch wenn durch die Neusortierung der Label (siehe Seite 79) ein gewisser Determinismus gegeben ist, ist ein direkter Vergleich zwischen den Labeln nicht angebracht. Es besteht kein semantischer Zusammenhang zwischen den o-Clustern der älteren und jüngeren Generation.

169 In diesem Fall die ältere Generation.

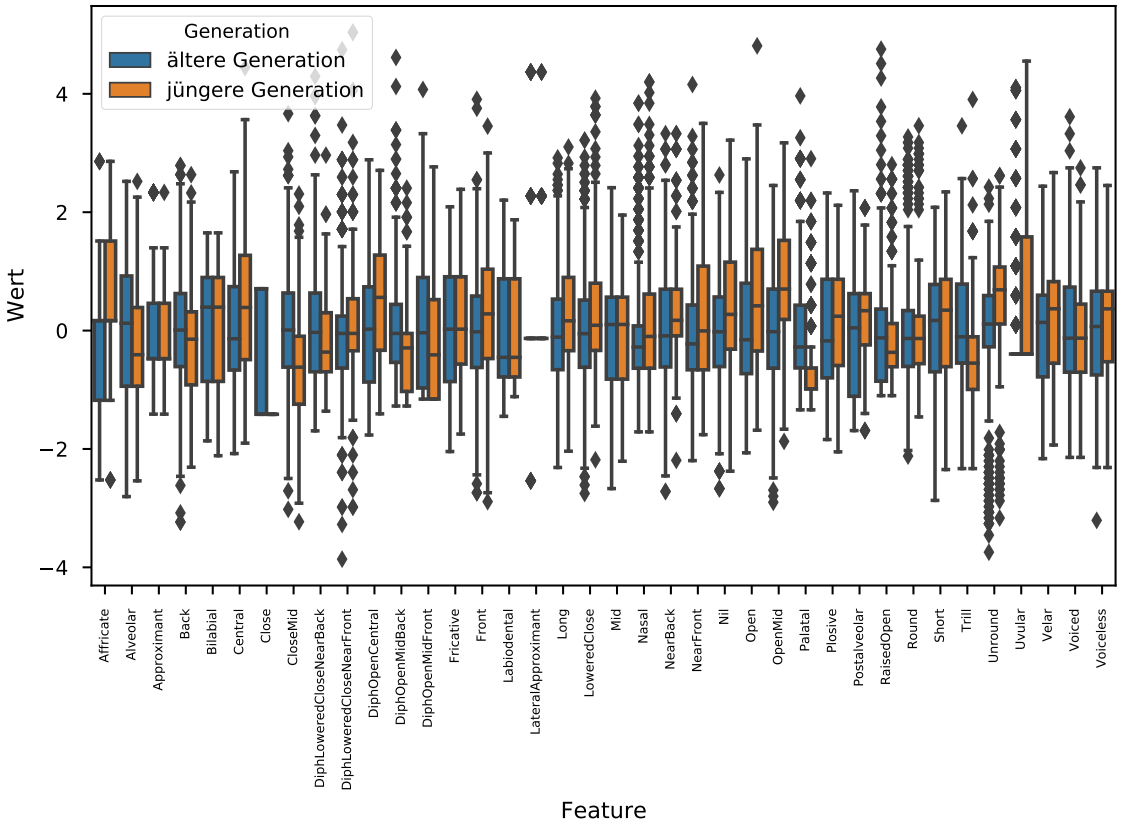



Abbildung 5.1: Vergleich der skalierten Verteilung des Datensets der älteren Generation und der jüngeren Generation.

Abbildung 5.1 zeigt eine Gegenüberstellung der Verteilungen der Datensets der älteren und der jüngeren Generation, wobei beide Generationen auf Basis der älteren Generation skaliert sind. Für viele Eigenschaften besteht eine deutliche Überlappung in der Verteilung zwischen den beiden Generationen. Auffällig ist, dass *Affricate* zwischen den Generationen gespiegelt ist. Dies weist auf einen eher überschaubaren Datenraum hin, der bei der Verteilung nicht viel Variation zulässt. So können bereits kleine Änderungen große Auswirkungen auf die Skalierung haben. Man kann auch erkennen, dass die Eigenschaft *Close* für die jüngere Generation völlig wegfällt[170]. Bei manchen Eigenschaften, wie *DiphOpenCentral* oder *OpenMid* verschiebt sich die Verteilung leicht ins Positive, wohingegen es sich bei *CloseMid* oder *DiphOpenMidBack* ins Negative verschiebt. Der Gegensatz zwischen den vielen positiven Ausreißern zu *Round* und den negativen zu *Unround* bleibt bestehen, woraus man schließen kann, dass zumindest das UMLAUTGEBIET als Sprachraum erhalten bleibt.

170 Bei der älteren Generation ist kein Medianbalken zu erkennen. Das bedeutet, dass der unskalierte Median 0 ist und damit bei der skalierten Verteilung am unteren Ende des Balkens liegen muss. Das *Close* der jüngeren Generation hat nur eine Markierung beim unteren Ende des Balkens der älteren Generation, damit müssen alle unskalierten Werte 0 sein.

## 5.2 ÄNDERUNGEN IN DEN CLUSTERN

Da die Daten einen Vektorraum aufspannen, lässt sich die euklidische Distanz zwischen skalierten Datenpunkten des Datensets der älteren Generation und den Datenpunkten der jüngeren Generation berechnen. Dies kann einen Einblick geben, in welchen Sprachräumen die größten Änderungen stattfinden. Abbildung 5.2 zeigt die Änderung in den in der Clusteranaylse bestimmten Hauptsprachräumen. Eine hellere Einfärbung der Orte symbolisiert eine größere Änderung. In Abbildung 5.3 sieht man die Reichweite der Änderungen nach den Labeln gruppiert. Insgesamt gibt es die größten Änderungen innerhalb des Umlautgebietes, die geringsten im Südpfälzischen Reliktgebiet. Das Hauptgebiet des Moselfränkischen hat die höchste Varianz bezüglich der Änderung. Räumlich betrachtet kann man also sagen, dass die Stärke der Änderung von Norden nach Süden abnimmt. Es ist wichtig zu beachten, dass man daraus keine Aussagen zur Änderung der Dialektalität treffen kann, da nur eine Differenz der älteren Generation berechnet wird, für eine Bewertung der Änderung der Dialektalität aber auch ein 0-Modell für die Standardsprache berücksichtigt werden muss. Allerdings kann eine geringe Änderung auf einen über die Zeit stabileren Sprachraum hindeuten[171].

Für einen genaueren Einblick, welche Lautklassen am stärksten von Änderungen betroffen sind, bietet es sich an, die Daten nach den historischen Bezugslauten zu gruppieren und die Differenz zwischen der älteren und der jüngeren Generation für diese Klassen zu berechnen. Abbildung 5.4 zeigt dies exemplarisch für die historischen Kurzvokale. Man sieht deutlich den starken Abbau von *CloseMid* und *Front* und eine auffallende Zunahme von *LoweredClose* und *NearFront* für *mhd. i* im 0-Cluster, also dem Cluster, das mit dem Umlautgebiet überlappt. Dies ist ein deutlicher Normalisierungsprozess von [e] → [ɪ] in dem ursprünglich sehr heterogenen Raum (siehe Abbildung 4.28, Seite 124). Die umgekehrte Richtung in *mhd. ë* allerdings fällt weitaus schwächer aus. Auch ist dieser Effekt nur im 0-Cluster für das Moselfränkische ausgeprägt. Das 1-Cluster, welches das übrige Gebiet der Reihenvertauschung einschließt, ist weniger stark von dieser Änderung betroffen. Auffällig ist ebenfalls die Zunahme der *Unround* Eigenschaft, besonders in den beiden erwähnten Lautklassen. Auch dies lässt einen Normalisierungsprozess vermuten.

Im 1-Cluster sieht man eine Rücknahme der Reihenvertauschung in *mhd. ü* und damit eine Annäherung an das Standarddeutsche. Während *Front* und *CloseMid* ([e]) abnehmen, nehmen *LoweredClose*, *NearFront* und *Round* ([ʏ]) zu. Im 2- und 3-Cluster fällt die Normalisierung von *mhd. a* auf. Es kommt zu einer Zunahme der [a̠] definierenden Eigenschaften, während die anderen Eigenschaften zurückgehen. Für das *mhd. i* reduzieren sich die Diphthongeigenschaften zu [e͡ɪ] (*DiphOpenMidFront* und *DiphLoweredCloseNearFront*) und die Monophthongeigenschaften zu [ɛ] und [ɪ] nehmen zu. Bei *mhd. ö* kommt es sogar noch deutlicher als im 1-Cluster zu einer Normalisierung in Richtung [ʏ]. Der Hauptunterschied bei den Änderungen in den Kurzvoka-

171 Ob diese Stabilität aus einer tief verwurzelten, generationenübergreifenden Dialektalität oder einer bereits bestehenden Nähe zum Standarddeutschen herrührt ist ohne zusätzliche Informationen nicht inferierbar.

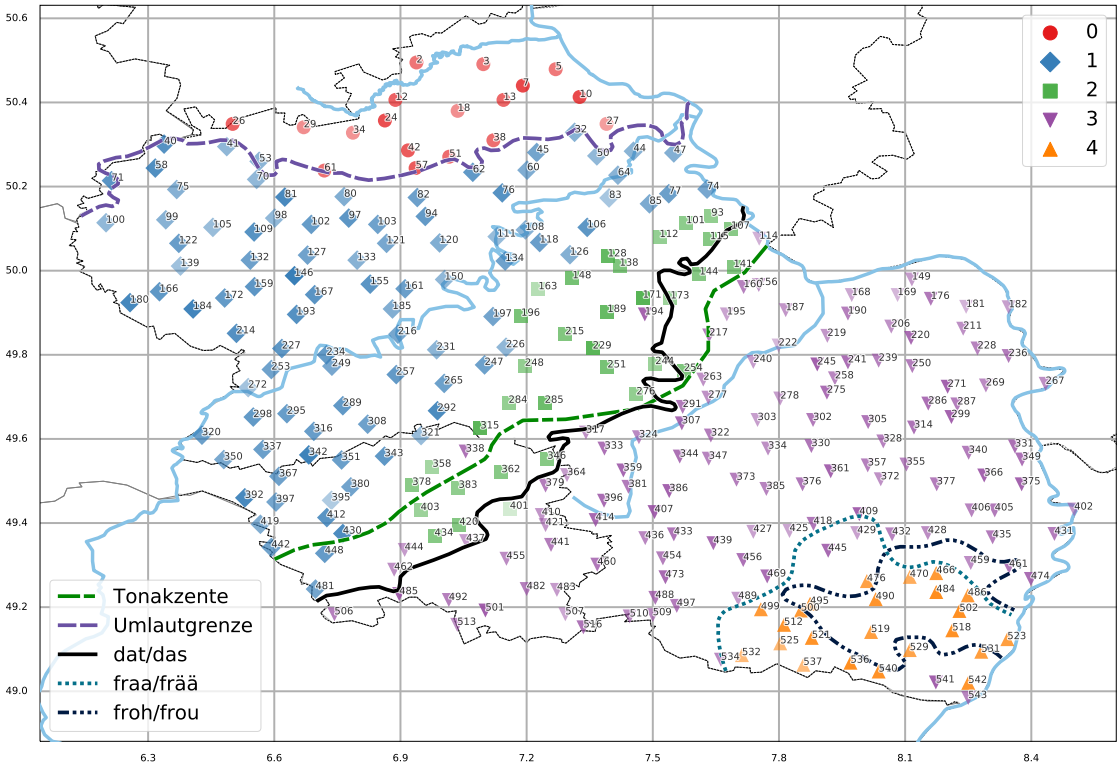



Abbildung 5.2: Änderung an den Datenpunkten zwischen der älteren und jüngeren Generation. Einfärbung der Orte nach dem WARD5-Clustering für alle Laute. Heller eingefärbte Orte haben eine größerer Änderung.

len zwischen dem 2- und 3-Cluster ist das breiter gestreute Vorkommen der Diphthonge und die leichte Zunahme im 3-Cluster, während die Diphthongeigenschaften im 2-Cluster rückläufig sind. Im 4-Cluster sind im Mittel die geringsten Änderungen zu beobachten. Die deutlichsten Änderungen finden sich in *mhd. e/ä* und *mhd. o* mit einer Anpassung in Richtung [e] beziehungsweise [o].

Die entsprechenden Grafiken zu den Langvokalen und den Konsonanten sind in Abschnitt A.6 (Seite 222) zu finden. Bei den Langvokalen gibt es im o-Cluster die größten Änderungen in *mhd. â* und *mhd. û*. Wobei im Ersteren die Eigenschaften zu [a͡o] deutlich zunehmen, während sich im Letzteren die durch die Reihendrehung beeinflussten Eigenschaften noch einmal verstärken. Auch gibt es wieder Normalisierungsprozesse bei *mhd. î*, allerdings weniger deutlich als bei den Kurzvokalen. Das 1- und 2-Cluster fallen in erster Linie durch den Mangel an signifikanten Änderungen[172] auf. Während es also durchaus zu Änderungen an den Lauteigenschaften kommt, befinden sich die meisten Änderungen zu den historischen Langvokalen in einem eher niedrigen Bereich. Außer in *mhd. â* und *mhd. æ* kommt es zu einer Abschwächung von *CloseMid* bei gleichzeitiger Stärkung von *OpenMid*.

172 Die mittlere Änderung, dargestellt in der ganz linken Spalte, zeigt die mittlere Änderung über alle Lauteigenschaften, also auch die kurzvokalischen. Dort sind auch die auffälligsten Änderungen zu verorten.

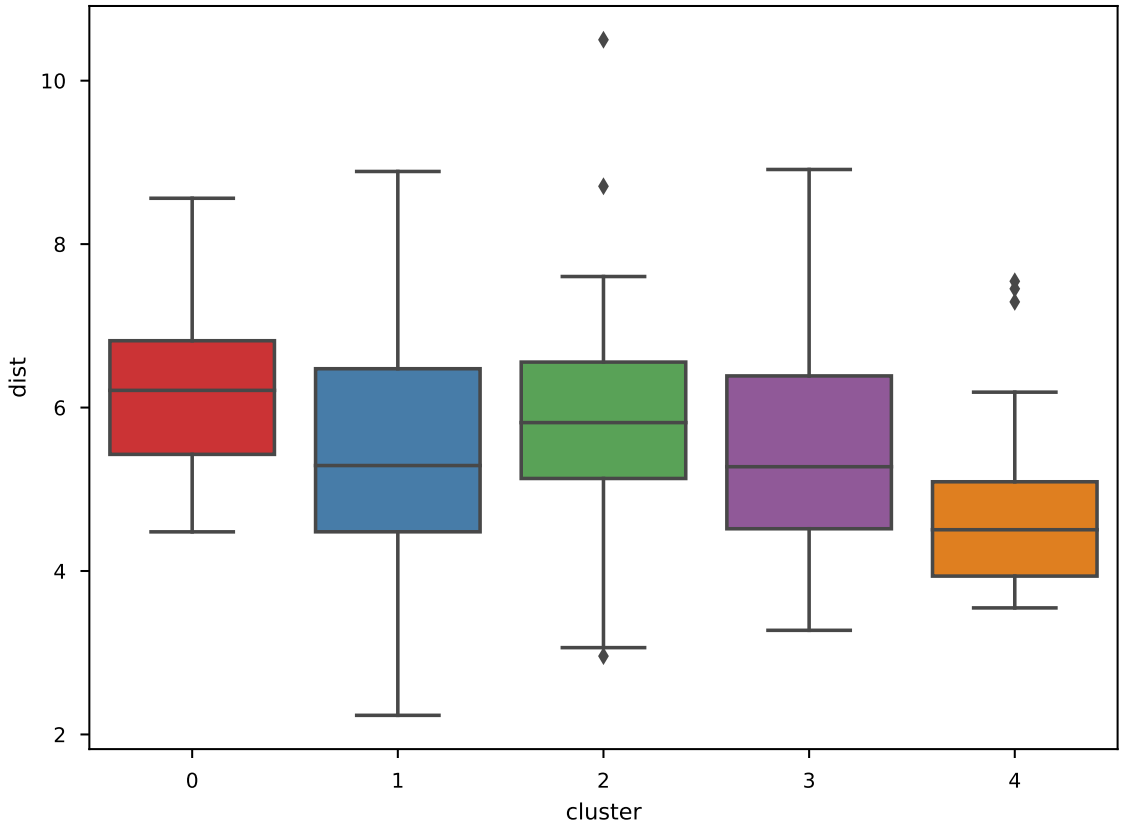



Abbildung 5.3: Spektrum der Änderungen in den Clustern nach WARD5 für alle Laute.

Auch im 3-Cluster gibt es die deutlichsten Änderungen in *mhd. â*. Diesmal mit einer merklichen Zunahme der [a̯] beeinflussenden Eigenschaften. Im 4-Custer findet vor allem ein Abbau der Diphthongeigenschaften bei *mhd. ô* und *mhd. œ* bei Zunahme entsprechender Monophthongeigenschaften (*Back* und *CloseMid* bei *mhd. ô* und *Front* und *OpenMid* bei *mhd. œ*) statt. Eine sehr deutliche Änderung gibt es noch bei *mhd. û*, indem die *DiphOpenCentral* gegen die *DiphOpenMidFront* Eigenschaft ausgetauscht wird.

Bei den Konsonanten fällt besonders *wg. r* auf. In dem 0- und 1-Cluster gibt es einen deutlichen Wechsel vom *alveolaren* [ɾ] hin zum *uvularen* [ʀ] oder Lautausfall (*Nil*). Im 2- und 3-Cluster verschiebt sich das *wg. r* in Richtung des vokalischen Schwa ([ə]) beziehungsweise Tiefschwa ([ɐ]) oder es fällt weg. In dem 4-Cluster nimmt vor allem der Ausfall des Lautes zu, während die anderen relevanten Lauteigenschaften für diese Klassen deutlich abnehmen. Insgesamt reduziert sich also das Lautspektrum von *wg. r*, wobei die Normalisierung zwischen den Clustern im Moselfränkischen und Rheinfränkischen nicht gleich verläuft. Bei allen konsonantischen Lautklassen kommt es zu einer deutlichen Zunahme des Ausfalls des Konsonanten, markiert durch die *Nil*-Eigenschaft.

**Box 5.2.1 Interpretationshilfe zu den Differenzgrafiken nach den historischen Lautklassen**

Grafiken wie Abbildung 5.4 zeigen die mittlere relative Änderung der Lauteigenschaften über alle Lautklassen und darüber hinaus aufgeteilt nach den historischen Lautklassen in jedem Custer. Die äußerst linke Spalte zeigt die Differenz gemittelt über alle Orte des Clusters und alle Lautklassen. Die anderen Spalten über die entsprechende Lautklasse und die Orte des Clusters.

Die Differenz ist ausgehend von den skalierten Daten berechnet, deswegen liegt der Wertebereich auch zwischen -3 und 3. Dabei sind alle Spalten unabhängig voneinander skaliert, wobei die ältere Generation jeweils als Ausgangsverteilung gesetzt ist.

Die Änderungen sind relativ zu sehen. Das bedeutet, wenn ein hypothetisches Feature einen Wert von -2 (sehr selten) in der älteren Generation hat und von 0 (mittlere Verteilung) in der jüngeren Generation, ist die Differenz genauso hoch, als wenn ein Feature 0 in der älteren Generation und 2 in der jüngeren hat. Ohne zusätzliche Informationen, lassen sich also nur schwer Rückschlüsse auf die Ausgangsverteilungen ziehen.

## 5.3 KLASSIFIKATION DER JÜNGEREN GENERATION

Wenn man das geclusterte Datenset der älteren Generation als Modell auffasst, kann man Klassifikatoren trainieren und diese gegen die jüngere Generation testen. Dies liefert einen Einblick, inwieweit die unterliegenden Sprachstrukturen über die Generation hinweg im Untersuchungsgebiet gewahrt wurden. Insgesamt ordnen Klassifikatoren basierend auf der älteren Generation die Datenpunkte der jüngeren Generation den entsprechenden Clustern zu. Es zeigt sich, dass ein Ort, der in der älteren Generation zum 2-Cluster gehört, auch mit einer hohen Genauigkeit in der jüngeren Generation denselben Clustern zugeordnet wird. Eine Dimensionseinbettung zeigt zudem, dass sich die Datenpunkte der jüngeren und älteren Generation überlappen. Es kommt also zu keiner signifikanten Änderung in der Datenstruktur. Dabei ist allerdings zu beachten, dass die Modelle in sich abgeschlossen sind. Ein Vergleich mit einem hypothetischen 0-Modell zur Standardsprache oder angrenzenden Sprachräumen findet nicht statt. Insofern können nur Änderungen innerhalb des Sprachraums gezeigt werden, und in dieser Hinsicht bleiben die Strukturen relativ zueinander weitestgehend erhalten.

Abbildung 5.5 zeigt eine Klassifizierung der jüngeren Generation auf Basis des WARD5 Models für alle Laute. Als Klassifikator kommt eine *Support Vector Machine* mit einem RBF Kernel[173] zum Einsatz. Nur wenige Orte haben

173 Eine *Support Vector Machine* (vgl. Cortes und Vapnik 1995; Cristianini und Shawe-Taylor 2000; Guyon, Boser und Vapnik 1993) ist einer der bekanntesten Klassifikationsalgorithmen und basiert auf der Unterteilung des Datenraums in Hyperebenen anhand von Stützvektoren. Stützvektoren sind die benachbarten Datenpunkte unterschiedlicher Klassen, deren Abstände zueinander es zu maximieren gilt und damit die (Vektorraum-)Basis der Trennebene

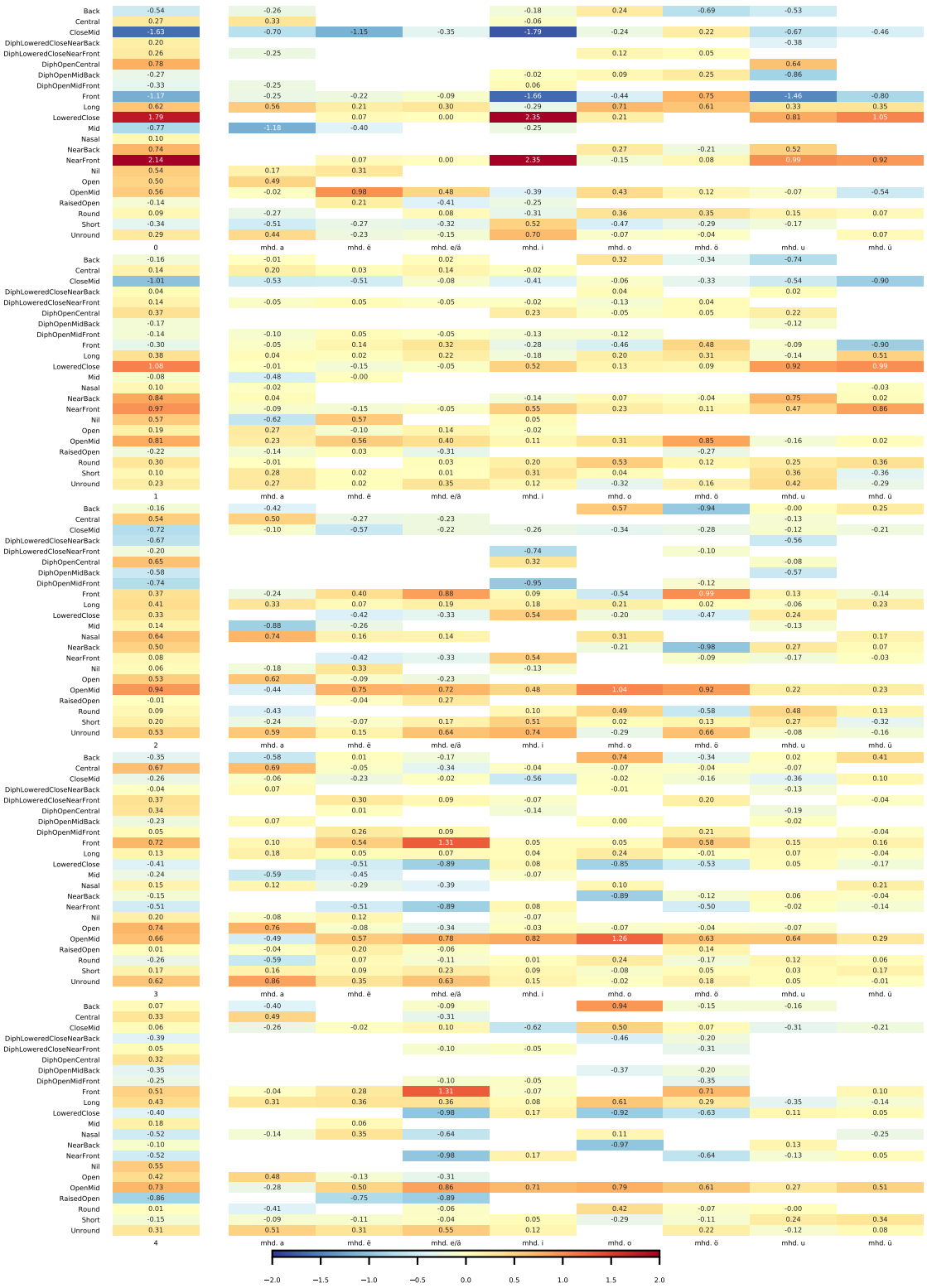



Abbildung 5.4: Spektrum der Änderungen in den Clustern nach WARD5 für die Lautklassen der historischen Kurzvokale.

stellen. Der *Radial Basis Function* Kernel (RBF) wird verwendet, um auch nicht linear separierbare Daten trennen zu können.

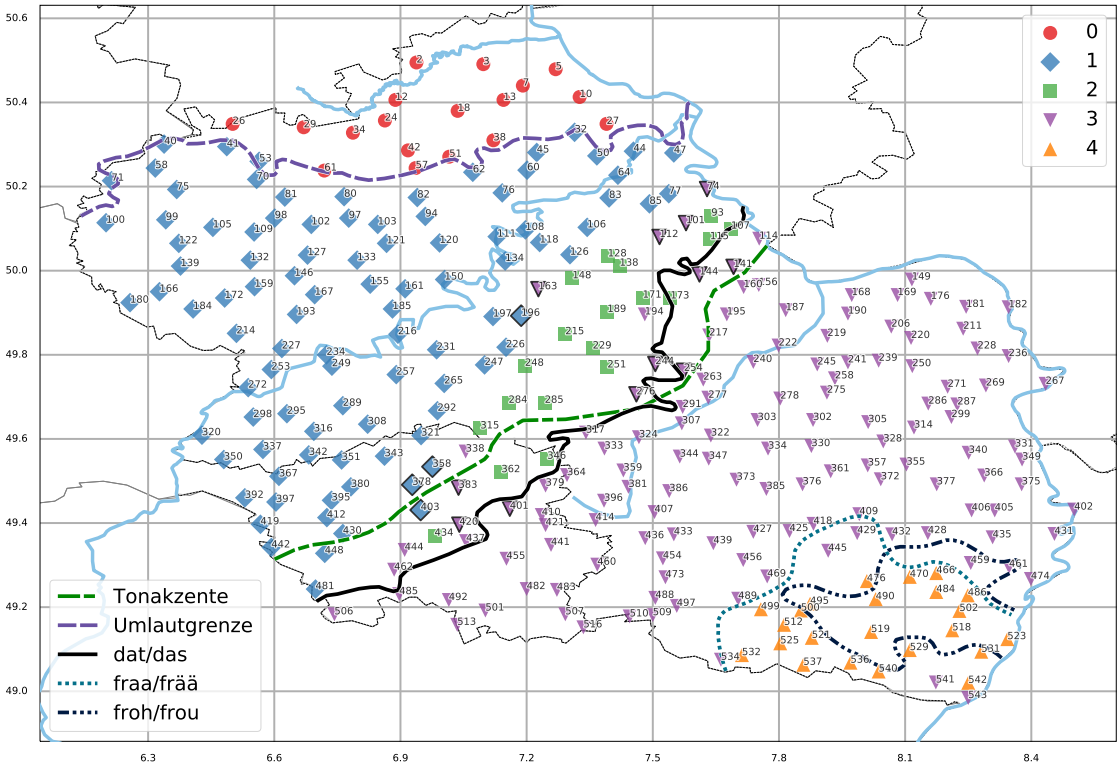



Abbildung 5.5: Klassifizierung der jüngeren Generation basierend auf dem WARD5 Clustering für alle Laute. Orte, denen ein anderes Label als in der älteren Generation zugewiesen wurde, sind schwarz umrandet.

eine schwarze Umrandung, die auf eine Änderung in der Labelzuweisung hinweist. Diese Stabilität lässt sich auch in den als gut befundenen Clusterings der anderen Experimente beobachten.

## 5.4 CLUSTERING DER JÜNGEREN GENERATION

Natürlich kann man dieselbe Methodik, die man zum Erstellen der Clusteranalysen der älteren Generation benutzt, auch auf die jüngere übertragen. Vergleiche zwischen den Generationen sind aber nur bedingt möglich, da diese Clusterings zum einen völlig unabhängig voneinander erstellt werden und zum anderen die Datenstruktur der jüngeren Generation zu einer anderen Clusterkonfiguration führen kann.

Abbildung 5.6 zeigt ein Clustering zur jüngeren Generation mit einem Silhouettenkoeffizienten von 0.18. Am auffälligsten ist bei diesem Clustering die Auflösung der scharfen Grenze entlang der *dat/das*-Grenze und der Tonakzentgrenze. Im 3-Clustering sind die vielen Orte mit negativer Silhouette prägnant, die sich besonders im südlichen Saarland häufen. Im Norden bleibt das 0-Cluster zum UMLAUTGEBIET stabil. Die Grenze zwischen dem 1- und 2-Cluster verläuft im nördlichen Bereich entlang der *Korf/Korb*-Grenze, fällt allerdings im südlichen Bereich nicht mit der *dat/das*-Grenze zusam-

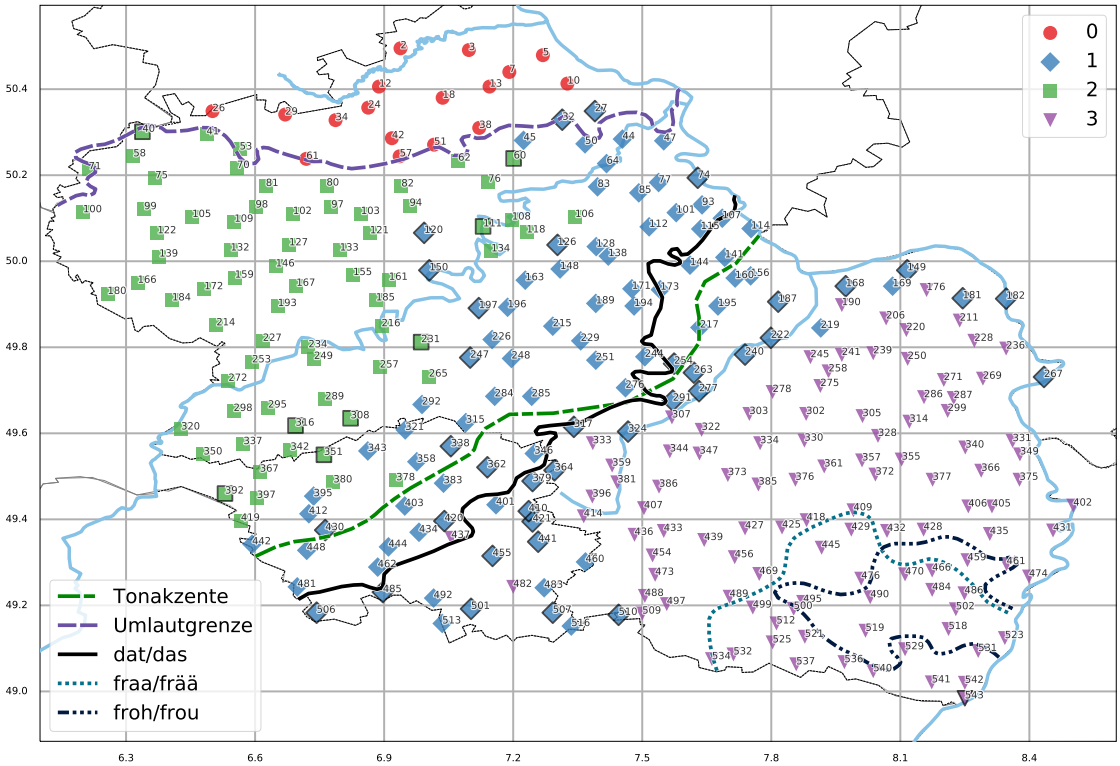



Abbildung 5.6: WARD4-Clustering auf allen Lauteigenschaften zur jüngeren Generation.

men, sondern bleibt eher in der Nähe der Tonakzentgrenze. Das 2-Cluster geht zudem über die *dat/das*-Isoglosse hinaus ins RHEINFRÄNKISCHE hinein und trennt sich erst im Norden in der Nähe vom 3-Cluster, im Süden verläuft die Grenze entlang der Grenze zwischen Saarland und Rheinland-Pfalz.

Die ANOVA (Abbildung 5.7) zeigt *Velar*, *Bilabial*, *Round* und *Labiodental* als einflussreichste Eigenschaften für die Struktur des Clusterings. Diese Eigenschaften weisen auch je einen der Hauptsprachräume auf. Auffällig ist die relativ niedrige Gewichtung von vokalischen Eigenschaften (außer Länge und Rundung). So ist *Central* die erste vokalische Eigenschaft. Diese Eigenschaft könnte mit Normalisierung in Richtung [a] einhergehen, die besonders im Gebiet des RHEINFRÄNKISCHEN auffällig ist.

Das Bootstrapping der Cluster (Abbildung 5.8) gibt eine Einsicht in die Stabilität der Cluster. Die Orte des 1-Clusters haben eine deutliche Nähe zum 0-Cluster, während die anderen Cluster etwas homogener sind. Es zeigt sich außerdem eine klare Übergangsregion. So beeinflusst das 1-Cluster das 2-Cluster vom Norden her, während das 3-Cluster das 2-Cluster vom Süden her beeinflusst. Durch diesen beidseitigen Einfluss kann das 2-Cluster als Übergangsregion interpretiert werden.

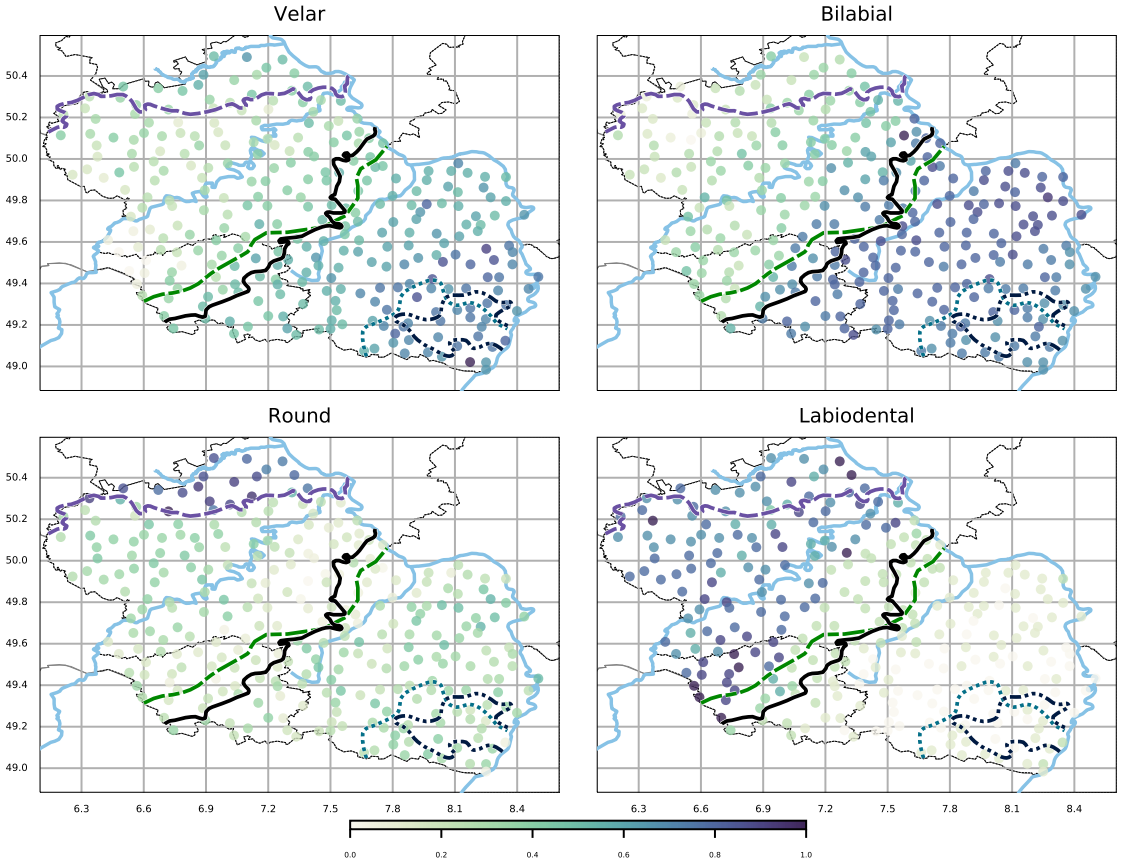



Abbildung 5.7: Räumliche Verteilung der vier einflussreichsten Features für WARD4 zu dem Datenset der jüngeren Generation.

## 5.5 ZUSAMMENFASSUNG

Die Struktur und Strukturgrenzen der älteren Generation bleiben auch in der jüngeren Generation weitgehend erhalten. Es kommt zwar zu Änderungen in den Dialekten, in vielen Fällen zu einer Angleichung an die Standardsprache, allerdings noch nicht in dem Maße, dass sich die Hauptsprachräume auflösen. In vielen Fällen sind Änderungen lautklassenbedingt und können sich je nach Sprachregion unterscheiden. So verschiebt sich das *alveolare* [r] im MOSELFRÄNKISCHEN in Richtung des *uvularen* [ʀ], wohingegen es im RHEINFRÄNKISCHEN eher wegfällt. Auch gibt es Bewegung in dem Gebiet der Reihenvertauschung. Während sich die ehemals sehr heterogenen Kurzvokale in Richtung des Standarddeutschen verändern, bleibt die Reihenvertauschung in den Langvokalen erhalten und verfestigt sich teilweise noch.

Insgesamt ist eine Abnahme der Änderungen von Norden (dem UMLAUTGEBIET) nach Süden (dem SÜDPFÄLZISCHEN RELIKTGEBIET) zu beobachten, wobei die Spannweite der Änderungen im Bereich des MOSELFRÄNKISCHEN am größten ist.

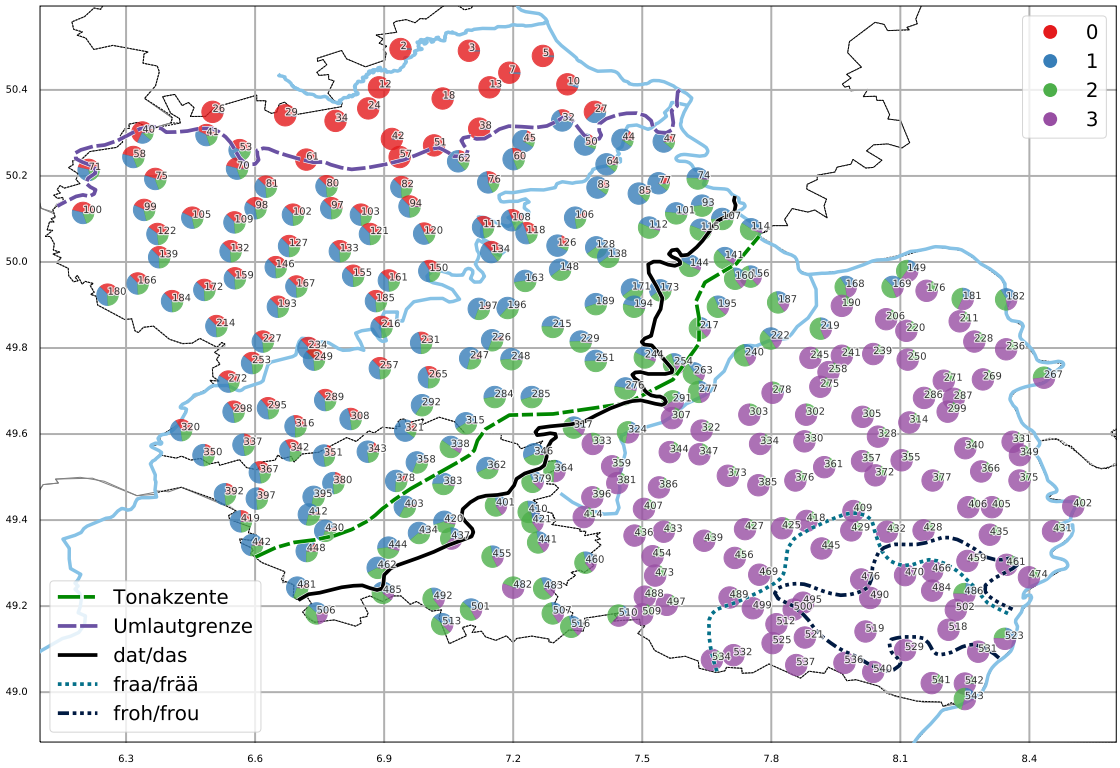



Abbildung 5.8: Bootstrapping auf KMEANS4-Clustering auf allen Lauteigenschaften zur jüngeren Generation.

# 6 ZUSAMMENFASSUNG UND AUSBLICK

Diese Arbeit stellt mit der *phonOntology* eine Möglichkeit vor, Laute, die in IPA annotiert sind, mittels Inferenz einer Menge an Lauteigenschaften zuzuordnen. So lässt sich ein [ɪː] als eine Kombination der ontologischen Klassen *Long, LoweredClose* und *NearFront* ausdrücken. Diese Transformation überführt zum einen IPA-Laute in ein für maschinelles Verarbeiten gut geeignetes Format und ermöglicht gleichzeitig eine differenzierte Sicht auf die Daten. Durch die Aufsplittung der Laute in ihre Eigenschaften lassen sich strukturelle Gemeinsamkeiten innerhalb der Daten untersuchen. Als Beispiel für eine derartige Untersuchung dient der *Mittelrheinische Sprachatlas*, da sich das verwendete, kontrollierte Vokabular auf IPA-Basis gut für eine solche Transformation eignet. Auf dem transformierten Datenset zu der älteren Generation werden verschiedene Clusteranalysen (Experimente) durchgeführt. Neben der Analyse zu allen Lauten werden auch die Laute zu den historischen Lang- und Kurzvokalen des Mittelhochdeutschen sowie den westgermanischen Konsonanten gesondert betrachtet. Es zeigt sich in allen Experimenten, dass es entlang der Tonakzentgrenze und der *dat/das*-Isoglosse eine Trennung in zwei Hauptcluster gibt. Diese Grenze ist sehr stabil und kann als die Hauptgrenze für diese Region betrachtet werden. Für die Langvokale findet die erste Trennung entlang der *wih/weh*-Isoglosse statt und für die Konsonanten entlang der *Korf/Korb*-Isoglosse. Die Hauptgrenze tritt bei höheren Clusterings ($k \geq 3$) zutage. Diese Abweichung von der Hauptgrenze zeigt zwei besondere Phänomene für diese Region und speziell für den Moselfränkischen Teil. Zum einen die sogenannte Reihenvertauschung, bei der die *Aperture*-Eigenschaften von /e/ – /o/ und /ɪ/ – /ʊ/ vertauscht sind, und zum anderen den *Labiodental/Fricative–Bilabial/Plosive*-Gegensatz, bei dem die Ausprägungshäufigkeiten gegensätzlich sind. Diese beiden Phänomene unterteilen den moselfränkischen Raum zusätzlich. Dies führt dazu, dass dieser Raum, der noch das sogenannte Umlautgebiet als Übergangsregion zum Ripuarischen umfasst, deutlich variantenreicher ist als das Rheinfränkische, bei dem im Süden das südpfälzische Reliktgebiet als signifikante Unterregion auftritt.

Die Tonakzente, die in den deutschen Dialekten ein exklusives Phänomen des Moselfränkischen sind, sind als eigenständige Eigenschaft ein stabilisierender Faktor bei den Experimenten, allerdings sind sie nicht ausschlaggebend für die Form der Cluster. Die Hauptgrenze findet sich auch in Experimenten, in denen die Tonakzente herausgefiltert wurden.

Ein Vergleich mit den Daten der jüngeren Generation zeigt, dass es zwar deutliche Normalisierungstendenzen[174] gibt, diese allerdings noch nicht ausreichen, um die Raumstrukturen neu zu ordnen oder aufzulösen. Die Hauptgrenze lässt sich in den Daten immer noch wiederfinden, allerdings sind die Unterschiede entlang der Grenze geringer, so dass sich ein breiteres Gebiet,

174 Ob diese Tendenzen eine Annäherung an das Standarddeutsch bedeuten ist zwar anzunehmen, kann aber anhand dieser Daten nicht inferiert werden, da es keine Referenzmenge gibt.

welches als Übergangsgebiet interpretiert werden kann, um die Tonakzentgrenze und die *dat/das*-Isoglosse findet.

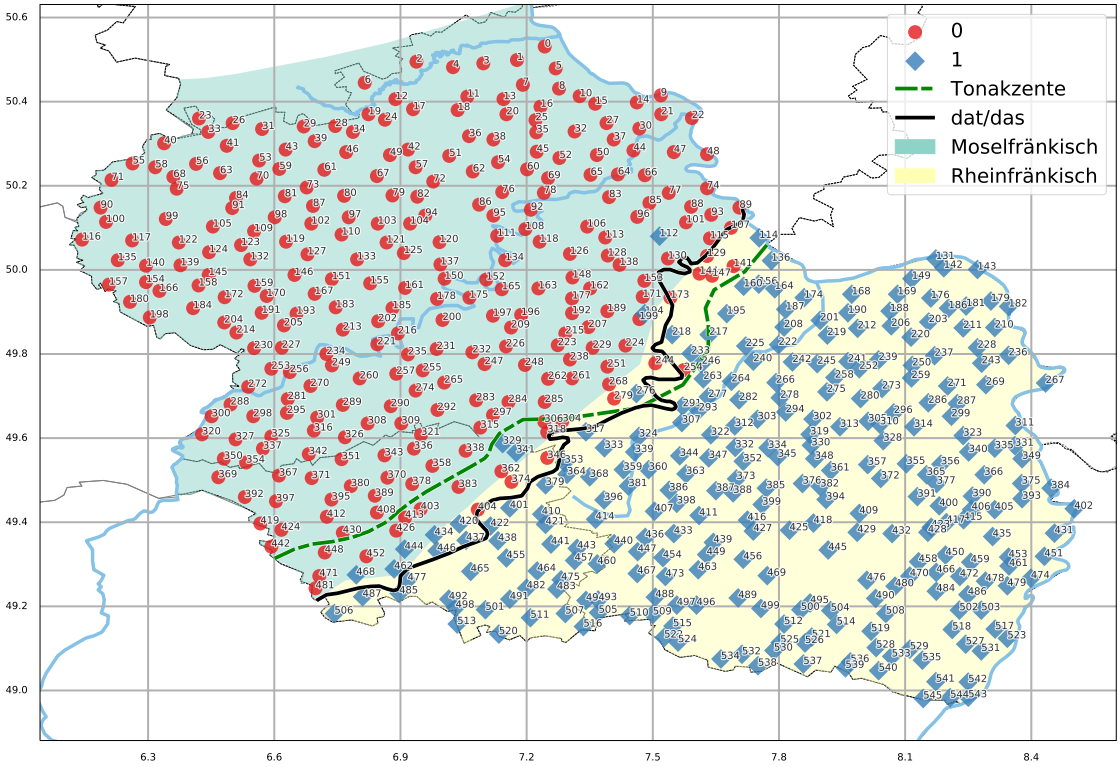



Abbildung 6.1: Vergleich eines KMEANS2-Clusterings zu allen Eigenschaften mit der Sprachraumeinteilung nach Wiesinger.

Strukturell bestätigt sich die Aufteilung nach Wiesinger (vgl. Wiesinger 1983, S. 830 f.) in das Moselfränkische und Rheinfränkische. Abbildung 6.1 zeigt den Vergleich mit den entsprechenden Raumstrukturen seiner Sprachraumstrukturierung. Der Adjusted-Rand-Index, bei dem die Einteilung nach Wiesinger als Ground Truth gesetzt wird, ist mit 0.87 sehr hoch[175].

## AUSBLICK

Eine Ontologie ist in der Informatik nicht nur ein Mittel, um Daten zu strukturieren und zu erweitern, sondern dient auch als Mediator zwischen verschiedenen Datenstrukturen. In diesem Sinne kann die *phonOntology* eingesetzt werden, um eine Kompatibilität zwischen verschiedenen digitalisierten Sprachatlanten herzustellen. Dieser Schritt würde weiterhin ein manuelles Mapping zwischen den Ausgangsdaten und der Ontologie benötigen, allerdings wären alle Daten, die über ein entsprechendes Mapping verfügen, untereinander vergleichbar. Damit kann die *phonOntology* als vereinheitlichendes Framework dienen, um eine Datengrundlage für eine übergreifen-

175 Es ist zu beachten, dass die Grenze in der Wiesingereinteilung eine vereinfachte Version der *dat/das*-Isoglosse ist und damit ein hoher ARI nicht verwunderlich ist.

de Analyse von Sprachatlanten zu generieren. Die *phonOntology* in der hier präsentierten Form repräsentiert nur einen Ausschnitt aller möglichen Laute und Lautvarianten, basierend auf den durch IPA definierten Lauten. Diakritika sind bis auf die Länge-Eigenschaft nicht weiter berücksichtigt. Da die Anzahl der möglichen Lautrepräsentationen durch das Hinzufügen von Diakritika natürlich deutlich ansteigt, auf Kosten der expliziten Benennung, gibt es noch deutliches Erweiterungsportenzial für die *phonOntology*. Als Ansatzpunkt kann hierbei das PHOIBLE-Datenset (vgl. Moran 2012; Moran, McCloy und Wright 2014) dienen, in dem eine Auflistung des Lautinventars verschiedener Sprachen weltweit, vorgenommen wird. Dabei muss allerdings beachtet werden, dass die Konstruktion von Lauteigenschaftskombinationen ontologisch gesehen kein größeres Problem darstellt, die explizite Zuordnung zu einem Laut kann jedoch an Grenzen stoßen. Es muss in vielen Fällen zwischen einer theoretisch korrekten und einer praktisch anwendbaren Repräsentation entschieden werden. Ein weiteres Problem, welches es bei einer Erweiterung der *phonOntology* durch das Einbeziehen von Diakritika zu lösen gilt, ist die Gewichtung der Laute. Viele Diakritika repräsentieren eine relative Änderung ausgehend von einer Lautkonfiguration. Dies wirft die Frage auf, ob eine Lauteigenschaft, die durch ein Diakritikum repräsentiert wird, dieselbe Gewichtung haben sollte, wie eine Haupteigenschaft. Diese Frage ist aber nicht an die Ontologie gebunden, sondern ein allgemeines linguistisches Forschungsfeld.

Mittels der *phonOntology* wurden Sprachraumstrukturierungen und dazugehörige Modelle für den MRhSA generiert. Es bietet sich natürlich an, weitere Modelle auf Basis der Ontologie zu erstellen, um zum Beispiel ein 0-Modell für das Standarddeutsch zu bekommen, das als Referenzmodell für Dialektalitätsbestimmungen dienen kann. Mit der von Herrgen und Schmidt entwickelten Methode zur Dialektalitätsmessung (vgl. Herrgen und Schmidt 1989) steht auch ein spezialisiertes Distanzmaß zur Differenzbestimmung zur Verfügung. Auch können die in dieser Arbeit generierten Modelle für Vergleiche mit anderen Sprachräumen herangezogen werden. So bietet der Audioatlas siebenbürgisch-sächsischer Dialekte[176] (ASD) (vgl. Krefeld, Lücke und Mages 2016) Material, das aufgearbeitet werden kann, um mit der *phonOntology* kompatibel zu werden. Indem ein Clustering zum Moselfränkischen als Klassifikationsmodell für ein kompatibles Datenset des ASD aufbereitet wird, kann geschaut werden, inwieweit sich der Dialekt der moselfränkischen Aussiedler in dieser Region bewahrt hat.

Eine weitere Anwendungsmöglichkeit ist eine Überprüfung oder Reindizierung der historischen Bezugssysteme. Die in dieser Arbeit verwendeten Bezugssysteme dienen in erster Linie als Referenzpunkt, um die Varianz eines Lautes in einem Wort, welches einem Bezugslaut zugeordnet ist, zu klassifizieren. Die *phonOntology* kann genutzt werden, um nach struktureller Ähnlichkeit von Wörtern zu suchen. So kann zum Beispiel nach räumlichen Zusammenhängen zwischen Lautrealisierungsverteilungen gesucht werden, um so Wörter zu finden, die einem ähnlichen Verteilungsmuster folgen, unabhängig ihrer zugeordneten historischen Klassifikation.

176 <http://doi.org/10.5282/asd>, abgerufen 15.10.2018.

Auch bieten sich noch weitere Analysemöglichkeiten für die Clusterings selbst. So werden in dieser Arbeit nur drei Algorithmen verwendet, die alle an ein vordefiniertes k gebunden sind. Die in Abschnitt 3.3 erwähnten dichtebasierten Clusterings können helfen, eine flexiblere Struktur in den Daten zu finden und im Zuge dessen neue, differenziertere Sichtweisen auf die Datenstruktur zu ermöglichen. Auch gibt es neben den drei hier vorgestellten Verfahren noch weitere Variationen von Dimensionseinbettung, die genutzt werden können, um Daten nicht als Cluster, sondern anhand eines stetigen Beschreibers im Raum abzubilden. Auch kann die Analyse der Features noch verfeinert werden. In *Detecting Shibboleths* (vgl. Prokić, Çöltekin und Nerbonne 2012) wird eine Methode vorgestellt, um sogenannte Shibboleths[177], also hervorstechende linguistische Charakteristika, innerhalb eines Clusters zu bestimmen. Ergebnisse solcher Analysen könnten wiederum in Sprachwahrnehmungsuntersuchungen verwendet werden.

Bei der Verwendung von semantischer Technologie bietet es sich immer an, diese auch in ein größeres Netzwerk einzubetten und als Onlinesystem zur Verfügung zu stellen. Webtechnologien ermöglichen eine einfache Erstellung angemessener Benutzeroberflächen, die dank REST[178]-API (vgl. Fielding und Taylor 2000; Fielding u. a. 2017) komplexe Serveranwendungen steuern können. Dies ist sehr hilfreich, um mehr moderne informatische Methoden und Prinzipien für die Dialektforschung anzubieten und zugänglich zu machen.

177 Engl.: distinctive, characteristic variants.
178 Representational State Transfer.

Teil III

# ANHANG

# A ANHANG

## A.1 PHONETISCHE EIGENSCHAFTEN DER GOLD ONTOLOGIE

- *ArticulatoryProperty*: The class of properties defining how sounds are produced in the mouth. (vgl. Ladefoged 1997)
  - *AirstreamProperty*: Refers to the direction of the airstream in speech sound production. In the canonical literature, there are three airstream mechanisms: pulmonic, velaric, and glottalic. Glottalic airstream mechanism is sometimes used to describe the method of production of ejectives and implosives. Ladefoged and Maddieson prefer to regard implosives and ejectives as characterized by a laryngeal parameter of movement rather than an airstream property. (vgl. Maddieson und Ladefoged 1996, S. 372–373)
    - *PulmonicProperty*: Pulmonic refers to an air-stream mechanism wherein the air is generated in the lungs and pushed out under the control of the respiratory muscles. (vgl. Ladefoged 2000, S. 122)
    - *VelaricProperty*: Velaric refers to an air-stream mechanism wherein the air is generated by a closure at the velar position, rather than an air-stream generated by the lungs. The back of the tongue is raised against the velum, and articulations are made farther forward by the lips or front parts of the tongue, drawing air into or pushing air out of the mouth. The clicks of some African languages are produced in this way. In English, they may be heard in the 'tut tut' sound. (vgl. Crystal 1985, S. 325–326; Hartmann 1973, S. 8)
  - *LaryngealProperty*: The laryngeal setting refers to differences in the timing of laryngeal activity in relation to oral articulation. Most languages have phonemic contrasts between classes of stops which differ in the mode of action of the larynx, or in the timing of laryngeal activity. (vgl. Maddieson und Ladefoged 1996, S. 47)
    - *GlottalMovementProperty*: A phonation type containing the features 'raising' and 'lowering'. (vgl. Maddieson und Ladefoged 1996, S. 372)
    - *GlottalStrictureProperty*: The three phonation types are part of the five possible values of Glottal Stricture that are used by languages. Sounds can have the vocal cords tightly together, as in a glottal stop, or they can be far apart as in voiceless sounds, or they can have one of the three phonation types: breathy voice, modal voice and creaky voice. Alt-

hough some phoneticians have shown how terms similar to these may be combinable from the phonetic point of view, the named terms form a set of phonologically mutually exclusive possibilities. These factors point to there being an ordered set of five possibilities: [voiceless], [breathy], [modal voice], [creaky] and [closed]. It is certainly appropriate to consider these glottal states as resulting from two physiological attributes of the vocal cords, their stiffness and their aperture. However from a linguistic point of view, the named values of the feature Glottal Stricture operate as a linearly ordered set of five mutually exclusive possibilities. (vgl. Ladefoged 1997, S. 607–608)

  * *GlottalTimingProperty*: A phonation type containing the features 'aspirated' and 'unaspirated. Aspiration involves matters of relativing timing between laryngeal and oral articulations, and the wider opening can be viewed as an aspect of the control of this timing. There are two ways of interpreting this greather width; it can be seen as the essential aspect of the production of voiceless aspiration, that is, aspiration is an extra-wide opening of the vocal folds [Kim 1965], or it can be seen as a by-product of the mechanism by which a delay between the offset of the oral and glottal gestures is achieved, that is, aspiration is esentially a matter of the timing between speech movements controlling laryngeal setting and oral articulation. (vgl. Maddieson und Ladefoged 1996; Ohala, Browman und Goldstein 1986, S. 49–66, 372)
  * *VoicingProperty*: Refers to the vibratory activity of the vocal folds. Most languages have phonemic contrasts between voiced and voiceless sounds (regular vibration of the vocal folds versus no vibration of the vocal folds respectively). However, Ladefoged and Maddieson recognize five steps in the continuum of modes of vibration in the glottis, going from breathy voice – the most open setting of the vocal folds in which vibration will occur, passing through slack voice, modal voice, and stiff voice, ending with creaky voice - the most constricted setting in which vibration will occur. Each of these modes of voicing may or may not be phonemic in a given language. (vgl. Maddieson und Ladefoged 1996, S. 48–49)

- *SupraLaryngealProperty*: The supralaryngeal node dominates the activity of all of the articulators except stiffening and slacking of the vocal folds. For consonants it can be viewed as the default node which comes into play when the supranasal node below it is deactivated. In the case of sounds produced by an articulator dominated by this node, the only possible segments are those which are traditionally classified as [-consonantal]. It is not necessary to specify manner features for sounds dominated by the

supralaryngeal node, because they are redundantly determined. (vgl. Keyser und Stevens 1994, S. 216)

* *MannerProperty*: A sound property referring to the kind of articulatory process used in a sound's production. The distinction between vowel and consonant is usually made in terms of manner of articulation. Within consonants, several articulatory types are recognized based on the type of closure made by the vocal organs. Within vowels, classification is based on the number of auditory qualities distinguishable in the sound, the position of the soft palate, and the type of lip position. (vgl. Crystal 1997, S. 232)
* *NasalityProperty*: The class of properties that describe the degree to which the velum or soft palate is raised or lowered, allowing or prohibiting air from escaping through the nose. (vgl. Kenstowicz 1993, S. 143)
* *PlaceProperty* The superclass of properties that specify the location of the articulators. (vgl. Ladefoged 1997, S. 594)

## A.2 LAUTDEFINITION DER *PHONONTOLOGY*

```
Prefix: : <http://issg.de/ontologies/phonetic#>
Prefix: dc: <http://purl.org/dc/elements/1.1/>
Prefix: owl: <http://www.w3.org/2002/07/owl#>
Prefix: rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Prefix: rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Prefix: xml: <http://www.w3.org/XML/1998/namespace>
Prefix: xsd: <http://www.w3.org/2001/XMLSchema#>
Ontology: <http://issg.de/ontologies/phonetic>

ObjectProperty: articulationManner
    SubPropertyOf:
        consonantProperty
    Domain:
        Consonant
    Range:
        ArticulationManner

ObjectProperty: articulationPhonation
    SubPropertyOf:
        consonantProperty
    Domain:
        Consonant
    Range:
        ArticulationPhonation

ObjectProperty: articulationPlace
    SubPropertyOf:
        consonantProperty
    Domain:
        Consonant
    Range:
        ArticulationPlace

ObjectProperty: consonantProperty
    SubPropertyOf:
        phoneticProperty
    Domain:
        Consonant
    Range:
        ConsonantArticulation

ObjectProperty: diphthongEnd
    Annotations:
        rdfs:comment "denotes the end position of a
        ↪  diphthong.",
        rdfs:label "diphthong end"
    SubPropertyOf:
        diphthongProperty
    Range:
        DiphthongEnd

ObjectProperty: diphthongProperty
    SubPropertyOf:
        vowelProperty
    Domain:
        Diphthong
    Range:
        DiphthongConfiguration
```

```
ObjectProperty: diphthongStart
    Annotations:
        rdfs:comment "denotes a starting position of a
        ↪  diphthong.",
        rdfs:label "diphthong start"
    SubPropertyOf:
        diphthongProperty
    Range:
        DiphthongStart

ObjectProperty: monophthongProperty
    SubPropertyOf:
        vowelProperty
    Domain:
        Monophthong
    Range:
        VowelConfiguration

ObjectProperty: phoneticProperty
    Domain:
        Phone
    Range:
        PhoneticProperty

ObjectProperty: prosodicProperty
    SubPropertyOf:
        phoneticProperty
    Domain:
        Phone
    Range:
        Intonation

ObjectProperty: vowelAperture
    SubPropertyOf:
        monophthongProperty
    Range:
        Aperture

ObjectProperty: vowelBackness
    SubPropertyOf:
        monophthongProperty
    Range:
        Backness

ObjectProperty: vowelLongness
    SubPropertyOf:
        monophthongProperty
    Range:
        Longness

ObjectProperty: vowelProperty
    SubPropertyOf:
        phoneticProperty
    Domain:
        Vowel

ObjectProperty: vowelRoundness
    SubPropertyOf:
        monophthongProperty
```

```
    Range:
        Roundness

Class: Affricate
    SubClassOf:
        ArticulationManner

Class: Alveolar
    SubClassOf:
        ArticulationPlace

Class: Aperture
    Annotations:
        rdfs:comment "Describes degree of openess during vowel
        ↪  creation"
    SubClassOf:
        VowelConfiguration

Class: Approximant
    SubClassOf:
        ArticulationManner

Class: ArticulationManner
    SubClassOf:
        ConsonantArticulation

Class: ArticulationPhonation
    SubClassOf:
        ConsonantArticulation

Class: ArticulationPlace
    SubClassOf:
        ConsonantArticulation

Class: Back
    SubClassOf:
        Backness

Class: Backness
    SubClassOf:
        VowelConfiguration

Class: Bilabial
    SubClassOf:
        ArticulationPlace

Class: Central
    SubClassOf:
        Backness

Class: Close
    SubClassOf:
        Aperture

Class: CloseBackRoundVowel
    Annotations:
        rdfs:label "u"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
```

```
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel

Class: CloseBackUnroundVowel
    Annotations:
        rdfs:label "ɯ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: CloseCentralRoundedVowel
    Annotations:
        rdfs:label "ʉ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Central)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: CloseCentralUnroundedVowel
    Annotations:
        rdfs:label "ɨ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Central)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: CloseFrontRoundedVowel
    Annotations:
        rdfs:label "y"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: CloseFrontUnroundedVowel
```

```
    Annotations:
        rdfs:label "i"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel
Class: CloseMid
    SubClassOf:
        Aperture

Class: CloseMidBackRoundedVowel
    Annotations:
        rdfs:label "o"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel

Class: CloseMidBackUnroundedVowel
    Annotations:
        rdfs:label "ɤ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: CloseMidCentralRoundedVowel
    Annotations:
        rdfs:label "ɵ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Central)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort

Class: CloseMidCentralUnroundedVowel
    Annotations:
        rdfs:label "ɘ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Central)
```

```
          and (vowelLongness value Short)
          and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort

Class: CloseMidFrontRoundedVowel
    Annotations:
        rdfs:label "ø"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: CloseMidFrontUnroundedVowel
    Annotations:
        rdfs:label "e"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel

Class: Consonant
    SubClassOf:
        Phone

Class: ConsonantArticulation
    SubClassOf:
        PhoneticProperty

Class: Dental
    SubClassOf:
        ArticulationPlace

Class: DiphLoweredCloseNearBack
    SubClassOf:
        DiphthongEnd

Class: DiphLoweredCloseNearFront
    SubClassOf:
        DiphthongEnd

Class: DiphOpenCentral
    SubClassOf:
        DiphthongStart

Class: DiphOpenCentralLoweredCloseNearBack
    Annotations:
        rdfs:label "a̯͡ʊ",
        rdfs:comment "The German a̯͡ʊ diphthong, following the
        ↪  convention at DSA."
```

```
    EquivalentTo:
        Vowel
         and ((diphthongEnd value DiphOpenCentral)
         and (diphthongStart value DiphLoweredCloseNearBack))
    SubClassOf:
        Diphthong

Class: DiphOpenCentralLoweredCloseNearFront
    Annotations:
        rdfs:label "a̟͡ɪ",
        rdfs:comment "The German a̟͡ʏ diphthong, following the
        ↪  convention at DSA."
    EquivalentTo:
        Vowel
         and ((diphthongEnd value DiphOpenCentral)
         and (diphthongStart value DiphLoweredCloseNearFront))
    SubClassOf:
        Diphthong

Class: DiphOpenFrontLoweredCloseNearBack
    Annotations:
        rdfs:label "ɛ͡ʊ"
    EquivalentTo:
        Vowel
         and ((diphthongEnd value DiphOpenMidFront)
         and (diphthongStart value DiphLoweredCloseNearBack))
    SubClassOf:
        Diphthong

Class: DiphOpenFrontLoweredCloseNearFront
    Annotations:
        rdfs:label "ɛ͡ɪ"
    EquivalentTo:
        Vowel
         and ((diphthongEnd value DiphOpenMidFront)
         and (diphthongStart value DiphLoweredCloseNearFront))
    SubClassOf:
        Diphthong

Class: DiphOpenMidBack
    SubClassOf:
        DiphthongStart

Class: DiphOpenMidFront
    SubClassOf:
        DiphthongStart

Class: DiphOpenMidLoweredCloseNearBack
    Annotations:
        rdfs:label "ɔ͡ʊ"
    EquivalentTo:
        Vowel
         and ((diphthongEnd value DiphOpenMidBack)
         and (diphthongStart value DiphLoweredCloseNearBack))
    SubClassOf:
        Diphthong

Class: DiphOpenMidLoweredCloseNearFront
    Annotations:
        rdfs:label "ɔ͡ɪ"
```

```
    EquivalentTo:
        Vowel
         and ((diphthongEnd value DiphOpenMidBack)
         and (diphthongStart value DiphLoweredCloseNearFront))
    SubClassOf:
        Diphthong

Class: Diphthong
    SubClassOf:
        Vowel

Class: DiphthongConfiguration
    SubClassOf:
        VowelConfiguration

Class: DiphthongEnd
    SubClassOf:
        DiphthongConfiguration

Class: DiphthongStart
    SubClassOf:
        DiphthongConfiguration
Class: ExtraShort
    SubClassOf:
        Longness

Class: Flap
    SubClassOf:
        ArticulationManner

Class: Fricative
    SubClassOf:
        ArticulationManner

Class: Front
    SubClassOf:
        Backness

Class: Gap
    Annotations:
        rdfs:label "-",
        rdfs:comment "Artifacial marker for a missing sound
        ↪  element"@en
    EquivalentTo:
        Phone
         and (phoneticProperty value Nil)
    SubClassOf:
        Phone

Class: Glottis
    SubClassOf:
        ArticulationPlace

Class: HalfLong
    SubClassOf:
        Longness

Class: Intonation
    Annotations:
        rdfs:label "Intonation"
```

```
    SubClassOf:
        PhoneticProperty

Class: Labiodental
    SubClassOf:
        ArticulationPlace
Class: LateralApproximant

    SubClassOf:
        ArticulationManner
Class: LateralFricative

    SubClassOf:
        ArticulationManner

Class: Long
    SubClassOf:
        Longness

Class: LongCloseBackRoundVowel
    Annotations:
        rdfs:label "uː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongCloseBackUnroundVowel
    Annotations:
        rdfs:label "ɯː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongCloseCentralRoundedVowel
    Annotations:
        rdfs:label "ʉː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Central)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongCloseCentralUnroundedVowel
    Annotations:
        rdfs:label "ɨː"
    EquivalentTo:
        Vowel
```

```
          and ((vowelAperture value Close)
          and (vowelBackness value Central)
          and (vowelLongness value Long)
          and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongCloseFrontRoundedVowel
    Annotations:
        rdfs:label "yː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongCloseFrontUnroundedVowel
    Annotations:
        rdfs:label "iː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Close)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongCloseMidBackRoundedVowel
    Annotations:
        rdfs:label "oː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongCloseMidBackUnroundedVowel
    Annotations:
        rdfs:label "ɤː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongCloseMidCentralRoundedVowel
    Annotations:
        rdfs:label "ɵː"
    EquivalentTo:
        Vowel
```

```
          and ((vowelAperture value CloseMid)
          and (vowelBackness value Central)
          and (vowelLongness value Long)
          and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongCloseMidCentralUnroundedVowel
    Annotations:
        rdfs:label "ɘː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Central)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongCloseMidFrontRoundedVowel
    Annotations:
        rdfs:label "øː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongCloseMidFrontUnroundedVowel
    Annotations:
        rdfs:label "eː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value CloseMid)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongLoweredCloseNearBackRoundedVowel
    Annotations:
        rdfs:label "ʊː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value LoweredClose)
         and (vowelBackness value NearBack)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongLoweredCloseNearFrontRoundedVowel
    Annotations:
        rdfs:label "ʏː"
    EquivalentTo:
        Vowel
```

```
          and ((vowelAperture value LoweredClose)
          and (vowelBackness value NearFront)
          and (vowelLongness value Long)
          and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongLoweredCloseNearFrontUnroundedVowel
    Annotations:
        rdfs:label "ɪː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value LoweredClose)
         and (vowelBackness value NearFront)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongMidCentralUnroundedVowe
    Annotations:
        rdfs:label "əː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Mid)
         and (vowelBackness value Central)
         and (vowelLongness value Long))
    SubClassOf:
        MonophthongLong

Class: LongOpenBackRoundedVowel
    Annotations:
        rdfs:label "ɒː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongOpenBackUnroundedVowel
    Annotations:
        rdfs:label "ɑː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongOpenCentralUnroundedVowel
        rdfs:label "a̟ː"
        rdfs:comment "The German long a"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
```

```
         and (vowelBackness value Central)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongOpenFrontRoundedVowel
    Annotations:
        rdfs:label "œː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongOpenFrontUnroundedVowel
    Annotations:
        rdfs:label "aː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongOpenMidBackRoundedVowel
    Annotations:
        rdfs:label "ɔː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongOpenMidBackUnroundedVowel
    Annotations:
        rdfs:label "ʌː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Back)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongOpenMidCentralRoundedVowel
    Annotations:
        rdfs:label "ɞː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
```

```
         and (vowelBackness value Central)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongOpenMidCentralUnroundedVowel
    Annotations:
        rdfs:label "ɜː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Central)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongOpenMidFrontRoundedVowel
    Annotations:
        rdfs:label "œː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongLong

Class: LongOpenMidFrontUnroundedVowel
    Annotations:
        rdfs:label "ɛː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Front)
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: LongRaisedOpenCentralVowel
    Annotations:
        rdfs:label "ɐː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value RaisedOpen)
         and (vowelBackness value Central)
         and (vowelLongness value Long))
    SubClassOf:
        MonophthongLong

Class: LongRaisedOpenFrontUnroundedVowel
    Annotations:
        rdfs:label "æː"
    EquivalentTo:
        Vowel
         and ((vowelAperture value RaisedOpen)
         and (vowelBackness value Front)
```

```
         and (vowelLongness value Long)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongLong

Class: Longness
    SubClassOf:
        VowelConfiguration

Class: LoweredClose
    SubClassOf:
        Aperture

Class: LoweredCloseNearBackRoundedVowel
    Annotations:
        rdfs:label "ʊ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value LoweredClose)
         and (vowelBackness value NearBack)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort

Class: LoweredCloseNearFrontRoundedVowel
    Annotations:
        rdfs:label "ʏ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value LoweredClose)
         and (vowelBackness value NearFront)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort

Class: LoweredCloseNearFrontUnroundedVowel
    Annotations:
        rdfs:label "ɪ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value LoweredClose)
         and (vowelBackness value NearFront)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort

Class: Mid
    SubClassOf:
        Aperture

Class: MidCentralUnroundedVowel
    Annotations:
        rdfs:label "ə"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Mid)
         and (vowelBackness value Central)
```

```
         and (vowelLongness value Short))
    SubClassOf:
        MonophthongShort

Class: Monophthong
    SubClassOf:
        Vowel

Class: MonophthongLong
    SubClassOf:
        Monophthong

Class: MonophthongShort
    SubClassOf:
        Monophthong

Class: Nasal
    SubClassOf:
        ArticulationManner

Class: NearBack
    SubClassOf:
        Backness

Class: NearFront
    SubClassOf:
        Backness

Class: Nil
    Annotations:
        rdfs:comment "Artifacial property for missing sound
        ↪  element (gap)"@en
    SubClassOf:
        PhoneticProperty
Class: Open
    SubClassOf:
        Aperture

Class: OpenBackRoundedVowel
    Annotations:
        rdfs:label "ɒ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: OpenBackUnroundedVowel
    Annotations:
        rdfs:label "ɑ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
```

```
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel

Class: OpenCentralUnroundedVowel
        rdfs:label "a̟",
        rdfs:comment "The German a"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Central)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort

Class: OpenFrontRoundedVowel
    Annotations:
        rdfs:label "Œ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort

Class: OpenFrontUnroundedVowel
    Annotations:
        rdfs:label "a"
    EquivalentTo:
        Vowel
         and ((vowelAperture value Open)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel

Class: OpenMid
    SubClassOf:
        Aperture

Class: OpenMidBackRoundedVowel
    Annotations:
        rdfs:label "ɔ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel

Class: OpenMidBackUnroundedVowel
    Annotations:
```

```
        rdfs:label "ʌ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Back)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: OpenMidCentralRoundedVowel
    Annotations:
        rdfs:label "ɵ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Central)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort

Class: OpenMidCentralUnroundedVowel
    Annotations:
        rdfs:label "ɜ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Central)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort

Class: OpenMidFrontRoundedVowel
    Annotations:
        rdfs:label "œ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Round))
    SubClassOf:
        MonophthongShort,
        SecondaryCardinalVowel

Class: OpenMidFrontUnroundedVowel
    Annotations:
        rdfs:label "ɛ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value OpenMid)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort,
        PrimaryCardinalVowel
```

```
Class: Palatal
    SubClassOf:
        ArticulationPlace

Class: Pharyngeal
    SubClassOf:
        ArticulationPlace

Class: Phone
    Annotations:
        rdfs:label "Phone"
Class: PhoneticProperty

Class: Plosive
    SubClassOf:
        ArticulationManner

Class: Postalveolar
    SubClassOf:
        ArticulationPlace

Class: PrimaryCardinalVowel
    SubClassOf:
        Vowel

Class: RaisedOpen
    SubClassOf:
        Aperture

Class: RaisedOpenCentralVowel
    Annotations:
        rdfs:label "ɐ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value RaisedOpen)
         and (vowelBackness value Central)
         and (vowelLongness value Short))
    SubClassOf:
        MonophthongShort

Class: RaisedOpenFrontUnroundedVowel
    Annotations:
        rdfs:label "æ"
    EquivalentTo:
        Vowel
         and ((vowelAperture value RaisedOpen)
         and (vowelBackness value Front)
         and (vowelLongness value Short)
         and (vowelRoundness value Unround))
    SubClassOf:
        MonophthongShort

Class: Retroflex
    SubClassOf:
        ArticulationPlace

Class: Round
    SubClassOf:
        Roundness
```

```
Class: Roundness
    SubClassOf:
        VowelConfiguration

Class: SecondaryCardinalVowel
    SubClassOf:
        Vowel

Class: Short
    SubClassOf:
        Longness

Class: TA1
    Annotations:
        rdfs:comment "Tonakzent 1 (nach Schmidt, Die
        ↪  Mittelfränkischen Tonakzente und MrhSA)"@de
    SubClassOf:
        Intonation

Class: TA2
    Annotations:
        rdfs:comment "Tonakzent 2 (nach Schmidt, Die
        ↪  Mittelfränkischen Tonakzente und MrhSA)"@de
    SubClassOf:
        Intonation

Class: Trill
    SubClassOf:
        ArticulationManner

Class: Unround
    SubClassOf:
        Roundness

Class: Uvular
    SubClassOf:
        ArticulationPlace

Class: Velar
    SubClassOf:
        ArticulationPlace

Class: Voiced
    SubClassOf:
        ArticulationPhonation

Class: VoicedAlveolarApproximant
    Annotations:
        rdfs:label "ɹ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Approximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarFricative
    Annotations:
```

```
        rdfs:label "z"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarLateralApproximant
    Annotations:
        rdfs:label "l"
    EquivalentTo:
        Consonant
         and ((articulationManner value LateralApproximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarLateralFricative
    Annotations:
        rdfs:label "ɮ"
    EquivalentTo:
        Consonant
         and ((articulationManner value LateralFricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarNasal
    Annotations:
        rdfs:label "n"
    EquivalentTo:
        Consonant
         and ((articulationManner value Nasal)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarPlosive
    Annotations:
        rdfs:label "t"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarTap
    Annotations:
        rdfs:label "ɾ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Flap)
         and (articulationPhonation value Voiced)
```

```
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedAlveolarTrill
    Annotations:
        rdfs:label "r"
    EquivalentTo:
        Consonant
         and ((articulationManner value Trill)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicedBilabialFricative
    Annotations:
        rdfs:label "β"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Bilabial))
    SubClassOf:
        Consonant

Class: VoicedBilabialNasal
    Annotations:
        rdfs:label "m"
    EquivalentTo:
        Consonant
         and ((articulationManner value Nasal)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Bilabial))
    SubClassOf:
        Consonant

Class: VoicedBilabialPlosive
    Annotations:
        rdfs:label "b"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Bilabial))
    SubClassOf:
        Consonant

Class: VoicedBilabialTrill
    Annotations:
        rdfs:label "в"
    EquivalentTo:
        Consonant
         and ((articulationManner value Trill)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Bilabial))
    SubClassOf:
        Consonant

Class: VoicedDentalFricative
```

```
    Annotations:
        rdfs:label "ð"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Dental))
    SubClassOf:
        Consonant

Class: VoicedGlottalFricative
    Annotations:
        rdfs:label "ɦ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Glottis))
    SubClassOf:
        Consonant

Class: VoicedLabiodentalApproximant
    Annotations:
        rdfs:label "ʋ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Approximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Labiodental))
    SubClassOf:
        Consonant

Class: VoicedLabiodentalFlap
    Annotations:
        rdfs:label "ⱱ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Flap)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Labiodental))
    SubClassOf:
        Consonant

Class: VoicedLabiodentalFricative
    Annotations:
        rdfs:label "v"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Labiodental))
    SubClassOf:
        Consonant

Class: VoicedLabiodentalNasal
    Annotations:
        rdfs:label "ɱ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Nasal)
```

```
         and (articulationPhonation value Voiced)
         and (articulationPlace value Labiodental))
    SubClassOf:
        Consonant

Class: VoicedPalatalApproximant
    Annotations:
        rdfs:label "j"
    EquivalentTo:
        Consonant
         and ((articulationManner value Approximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicedPalatalFricative
    Annotations:
        rdfs:label "ʝ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicedPalatalLateralApproximant
    Annotations:
        rdfs:label "ʎ"
    EquivalentTo:
        Consonant
         and ((articulationManner value LateralApproximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicedPalatalNasal
        rdfs:label "ɲ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Nasal)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicedPalatalPlosive
    Annotations:
        rdfs:label "ɟ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicedPharyngealFricative
```

```
    Annotations:
        rdfs:label "ʕ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Pharyngeal))
    SubClassOf:
        Consonant

Class: VoicedPostalveolarAffricate
    Annotations:
        rdfs:label "d͡ʒ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Affricate)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Postalveolar))
    SubClassOf:
        Consonant

Class: VoicedPostalveolarFricative
    Annotations:
        rdfs:label "ʒ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Postalveolar))
    SubClassOf:
        Consonant

Class: VoicedRetroflexApproximant
    Annotations:
        rdfs:label "ɻ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Approximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicedRetroflexFricative
    Annotations:
        rdfs:label "ʐ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicedRetroflexLateralApproximant
    Annotations:
        rdfs:label "ɭ"
    EquivalentTo:
        Consonant
         and ((articulationManner value LateralApproximant)
```

```
         and (articulationPhonation value Voiced)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicedRetroflexNasal
    Annotations:
        rdfs:label "ɳ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Nasal)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicedRetroflexPlosive
    Annotations:
        rdfs:label "ɖ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicedRetroflexTap
    Annotations:
        rdfs:label "ɾ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Flap)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicedUvularFricative
    Annotations:
        rdfs:label "ʁ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Uvular))
    SubClassOf:
        Consonant

Class: VoicedUvularNasal
    Annotations:
        rdfs:label "ɴ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Nasal)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Uvular))
    SubClassOf:
        Consonant
```

```
Class: VoicedUvularPlosive
    Annotations:
        rdfs:label "ɢ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Uvular))
    SubClassOf:
        Consonant

Class: VoicedUvularTrill
    Annotations:
        rdfs:label "ʀ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Trill)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Uvular))
    SubClassOf:
        Consonant

Class: VoicedVelarApproximant
    Annotations:
        rdfs:label "ɰ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Approximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: VoicedVelarFricative
    Annotations:
        rdfs:label "ɣ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: VoicedVelarLateralApproximant
    Annotations:
        rdfs:label "ʟ"
    EquivalentTo:
        Consonant
         and ((articulationManner value LateralApproximant)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: VoicedVelarNasal
    Annotations:
        rdfs:label "ŋ"
    EquivalentTo:
        Consonant
```

```
         and ((articulationManner value Nasal)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: VoicedVelarPlosive
    Annotations:
        rdfs:label "g"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiced)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: Voiceless
    SubClassOf:
        ArticulationPhonation

Class: VoicelessAlveolarAffricate
    Annotations:
        rdfs:label "t͡s"
    EquivalentTo:
        Consonant
         and ((articulationManner value Affricate)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicelessAlveolarFricative
    Annotations:
        rdfs:label "s"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicelessAlveolarLateralFricative
    Annotations:
        rdfs:label "ɬ"
    EquivalentTo:
        Consonant
         and ((articulationManner value LateralFricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicelessAlveolarPlosive
    Annotations:
        rdfs:label "d"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
```

```
         and (articulationPhonation value Voiced)
         and (articulationPlace value Alveolar))
    SubClassOf:
        Consonant

Class: VoicelessBilabialFricative
    Annotations:
        rdfs:label "ɸ̟"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Bilabial))
    SubClassOf:
        Consonant

Class: VoicelessBilabialLabiodentalAffricate
    Annotations:
        rdfs:label "p͡f"
    EquivalentTo:
        Consonant
         and (((articulationPlace value Bilabial)
         and (articulationPlace value Labiodental))
         and (articulationManner value Affricate)
         and (articulationPhonation value Voiceless))
    SubClassOf:
        Consonant

Class: VoicelessBilabialPlosive
    Annotations:
        rdfs:label "p"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Bilabial))
    SubClassOf:
        Consonant

Class: VoicelessDentalFricative
    Annotations:
        rdfs:label "θ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Dental))
    SubClassOf:
        Consonant

Class: VoicelessGlottalFricative
    Annotations:
        rdfs:label "h"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Glottis))
    SubClassOf:
        Consonant
```

```
Class: VoicelessGlottalPlosive
    Annotations:
        rdfs:label "ʔ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Glottis))
    SubClassOf:
        Consonant

Class: VoicelessLabiodentalFricative
    Annotations:
        rdfs:label "f"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Labiodental))
    SubClassOf:
        Consonant

Class: VoicelessPalatalFricative
    Annotations:
        rdfs:label "ç"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicelessPalatalPlosive
    Annotations:
        rdfs:label "c"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Palatal))
    SubClassOf:
        Consonant

Class: VoicelessPharyngealFricative
    Annotations:
        rdfs:label "ħ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Pharyngeal))
    SubClassOf:
        Consonant

Class: VoicelessPostalveolarAffricate
    Annotations:
        rdfs:label "t͡ʃ"
    EquivalentTo:
```

```
        Consonant
         and ((articulationManner value Affricate)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Postalveolar))
    SubClassOf:
        Consonant

Class: VoicelessPostalveolarFricative
    Annotations:
        rdfs:label "ʃ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Postalveolar))
    SubClassOf:
        Consonant

Class: VoicelessRetroflexFricative
    Annotations:
        rdfs:label "ʂ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicelessRetroflexPlosive
    Annotations:
        rdfs:label "ʈ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Retroflex))
    SubClassOf:
        Consonant

Class: VoicelessUvularFricative
    Annotations:
        rdfs:label "χ"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Uvular))
    SubClassOf:
        Consonant

Class: VoicelessUvularPlosive
    Annotations:
        rdfs:label "q"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Uvular))
    SubClassOf:
```

```
        Consonant

Class: VoicelessVelarAffricate
    Annotations:
        rdfs:label "k͡x"
    EquivalentTo:
        Consonant
         and ((articulationManner value Affricate)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: VoicelessVelarFricative
    Annotations:
        rdfs:label "x"
    EquivalentTo:
        Consonant
         and ((articulationManner value Fricative)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: VoicelessVelarPlosive
    Annotations:
        rdfs:label "k"
    EquivalentTo:
        Consonant
         and ((articulationManner value Plosive)
         and (articulationPhonation value Voiceless)
         and (articulationPlace value Velar))
    SubClassOf:
        Consonant

Class: Vowel
    SubClassOf:
        Phone

Class: VowelConfiguration
    SubClassOf:
        PhoneticProperty

Individual: Affricate
    Types:
        ArticulationManner

Individual: Alveolar
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: Approximant
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: Back
```

```
    Types:
        Backness,
        PhoneticProperty,
        VowelConfiguration

Individual: Bilabial
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: Central
    Types:
        Backness,
        PhoneticProperty,
        VowelConfiguration

Individual: Close
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: CloseMid
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: Dental
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: DiphLoweredCloseNearBack
    Types:
        DiphthongConfiguration,
        DiphthongEnd,
        PhoneticProperty,
        VowelConfiguration

Individual: DiphLoweredCloseNearFront
    Types:
        DiphthongConfiguration,
        DiphthongEnd,
        PhoneticProperty,
        VowelConfiguration

Individual: DiphOpenCentral
    Types:
        DiphthongConfiguration,
        DiphthongStart,
        PhoneticProperty,
        VowelConfiguration

Individual: DiphOpenMidBack
    Types:
        DiphthongConfiguration,
        DiphthongStart,
```

```
        PhoneticProperty,
        VowelConfiguration

Individual: DiphOpenMidFront
    Types:
        DiphthongConfiguration,
        DiphthongStart,
        PhoneticProperty,
        VowelConfiguration

Individual: ExtraShort
    Types:
        Longness,
        PhoneticProperty,
        VowelConfiguration

Individual: Flap
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: Fricative
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: Front
    Types:
        Backness,
        PhoneticProperty,
        VowelConfiguration

Individual: Glottis
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: HalfLong
    Types:
        Longness,
        PhoneticProperty,
        VowelConfiguration

Individual: Labiodental
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: LateralApproximant
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: LateralFricative
    Types:
```

```
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: Long
    Types:
        Longness,
        PhoneticProperty,
        VowelConfiguration

Individual: LoweredClose
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: Mid
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: Nasal
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: NearBack
    Types:
        Backness,
        PhoneticProperty,
        VowelConfiguration

Individual: NearFront
    Types:
        Backness,
        PhoneticProperty,
        VowelConfiguration

Individual: Nil
    Types:
        PhoneticProperty

Individual: Open
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: OpenMid
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: Palatal
    Types:
        ArticulationPlace,
        ConsonantArticulation,
```

```
        PhoneticProperty

Individual: Pharyngeal
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: Plosive
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: Postalveolar
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: RaisedOpen
    Types:
        Aperture,
        PhoneticProperty,
        VowelConfiguration

Individual: Retroflex
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: Round
    Types:
        PhoneticProperty,
        Roundness,
        VowelConfiguration

Individual: Short
    Types:
        Longness,
        PhoneticProperty,
        VowelConfiguration

Individual: TA1
    Types:
        PhoneticProperty
Individual: TA2
    Types:
        PhoneticProperty

Individual: Trill
    Types:
        ArticulationManner,
        ConsonantArticulation,
        PhoneticProperty

Individual: Unround
    Types:
        PhoneticProperty,
```

```
        Roundness,
        VowelConfiguration

Individual: Uvular
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: Velar
    Types:
        ArticulationPlace,
        ConsonantArticulation,
        PhoneticProperty

Individual: Voiced
    Types:
        ConsonantArticulation,
        PhoneticProperty,
        Voiced

Individual: Voiceless
    Types:
        ConsonantArticulation,
        PhoneticProperty,
        Voiceless
```

## A.3 BEZUGSLAUTE UND REFERENZWORTE

Tabelle A.1: Das Bezugssystem des MRhSA und zugehörige Referenzlaute.

| BEZUGSLAUT | REFERENZWORT |
|---|---|
| mhd. a | Asche, sagen, das, an, Mann, Nase, Schatten, Tag, Montag, Nagel, Hammer, wann, Arbeit, Garten, allein, Papier |
| mhd. e/ä | Zähne, Blätter, Kälte, steck, Egge, Nägel, Hefe, dreschen, fremde, wenn, Hölle, stellen, gerben |
| mhd. ei | Ei, zweite, Kleider, Fleisch, kein, ein, einen, weiß, Krankheit, Kleid, leid, Eidechse, Teig, Meister, Eimer, zwanzig |
| mhd. i | isst, nichts, gebissen, Kirche, mit, Kirsche, ist, ich, nicht, Kind, Milch, Honig, sieben, Giebel, Schlitten, Fisch, Vieh, Trichter, schwimmen, Stiel, Birne |
| mhd. ie | die, vier, lieb, fliegen, Brief, passieren, Knie, Lied, Spiegel, Licht, Riemen, Papier |
| mhd. iu | Leute, Säufer, neue, Feuer, heute, euch, neun, eure, Läuse, teuer, Häute, feucht |
| mhd. o | Ochse, Boden, Vogel, trocken, Ofen, gebrochen, Kohlen, vor, gestorben, Frosch, wohnen, Honig, voll, Boden, Ofen, Honig, Handvoll |
| mhd. ou | Auge, taub, Frau, behauen, glauben, auch, Baum |
| mhd. u | Hunde, du, Durst, durch, Hund, Zucker, Kugel, Lust, Frucht, Fuchs, herum, Pfund, schuld, schuldig, Lust, Fuchs, Pfund |
| mhd. uo | zu, Mutter, muß, tun, Hut, Futter, Fuß, Huf, rufst, Schuh, Handschuh, suchen, Blume, Stuhl, Schule, zu, Hut, Fuß |
| mhd. â | gelassen, blau, Montag, Adern, gebracht, nach, wo, schlafen, getan, schlagen, ohne, Nachbar, Samen, Naht, blau, Naht, getan |
| mhd. æ | läßt, täte, Schäfer, nähen, spät, sät, schläft, Käse, schwer |
| mhd. ê | weh, stehen, erste, Zeh, Peter |
| mhd. ë | Herde, er, geben, Mehl, Herd, herum, Wetter, Pfeffer, gestern, dem, Berg, Stecken, Regen, Nest, Fenster, hell, Wetter, Pfeffer, Fenster, (v)erzählt |
| mhd. î | gleich, hinein, bei, bleib, Zeiten, weiß, Eis, Wein, Feiertagen, Hochzeit, Seide, Scherenschleifer, Deichsel |
| mhd. ô | Hochzeit, rot, Ohren, roh, froh, Floh, Rose, Bohne, froh, rot, Rose |
| mhd. ö | Vögel, Köpfe, Frösche, Öl, Körbe, Hörner |

Tabelle A.1: Das Bezugssystem des MRhSA und zugehörige Referenzlaute.

| BEZUGSLAUT | REFERENZWORT |
|---|---|
| mhd. öu | Bäume, Heu, freuen, streuen, Krauthäuptchen, läuft, träumen |
| mhd. û | Raupe, bauen, Bauern, aufräumen, Haus, heraus, Braut, Daumen, Zaun, Uhren, Braut, Raupe, Zaun, Nachbar |
| mhd. ü | Stückchen, fünf, Schlüssel, Küche, kommst, Bühne, Mühle, Tür, dürr, Gürtel, Schürze, Stückchen, Schlüssel, Küche |
| mhd. üe | Hüte, Füße, müssen, Pflüge, Bücher, grün, Stühle, rühren |
| mhd. œ | böse, schöne, hört, Flöhe, Röteln, schöner |
| wg. b | gibt, Arbeit, ab, gestorben, bleib, Abend, Körbe, Gabel, geben, Taube, Kamm, taub |
| wg. d | Leute, Blätter, Montag, alte, Pfund, Futter, Beutel, Schulter |
| wg. f | Luft, fünfzig, Hof, voll, Armvoll, Hefe, Ofen |
| wg. g | Kugel, grün, Tag, Garten, fliegen, morgen, Berg, Montag, Pflüge, Pflug, Waage, Wagen, Nagel, Orgel |
| wg. h | höher, Nachbar, Frucht, hoch, nichts, Nacht, Ochse, wachsen, nach, Deichsel, Flöhe, Licht, Schuh, Floh, Hochzeit |
| wg. j | sät, jünger |
| wg. k | ich, gebacken, welcher, Kind, Kleid, trocken, macht, auch, schmeckt, suchen, Seiche, Kirche, Markt |
| wg. l | hell, kalt, mahlen |
| wg. m | kommt, kommst, Besen, Armvoll, Faden |
| wg. n | lang, an, seine, schöner, Abend, Hund, uns, wachsen, getan, ein, Wein, neun, Ast, Stühlchen, Zange, Fenster, Hörner, schreiben, Bein, Zaun |
| wg. p | aufräumen, Pfund, Apfel, gekauft, Pfennig, Kupfer, Knopf, pfeifen, Stiefkind, scharf, Spiegel |
| wg. r | Berg, Butter, dürr, gestorben, Garten, Durst, Rose, Haare, rühren, fahren, Kirsche, Kirche, Birne, vier |
| wg. s | waschen, festklopfen, Salz, Besen, ist, unser, erste |
| wg. t | kurz, eins, was, zu, müssen, beißen, es, Nacht, Weizen, gesetzt, gelassen, bis auf, bitter, Markt, Kirsche |
| wg. w | Farbe, was, Zehen |
| wg. þ | Kleider, Zeiten, mit, Hemd, schneiden, fremde, Egge, Zahn |

## A.4 ORTE DES MITTELRHEINISCHEN SPRACHATLAS

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 0 | 99650 | Koisdorf |
| 1 | 94716 | Königsfeld |
| 2 | 97570 | Lind |
| 3 | 100320 | Ramersbach |
| 4 | 100132 | Kesseling |
| 5 | 95181 | Gönnersdorf |
| 6 | 98404 | Wershofen |
| 7 | 97418 | Galenberg |
| 8 | 98860 | Wassenach |
| 9 | 96714 | Urmitz |
| 10 | 96621 | Nickenich |
| 11 | 96260 | Hohenleimbach |
| 12 | 100794 | Reifferscheid |
| 13 | 94747 | Weibern |
| 14 | 99858 | Kettig |
| 15 | 97789 | Kretz |
| 16 | 95285 | Bell |
| 17 | 96118 | Adenau |
| 18 | 96433 | Siebenbach |
| 19 | 97387 | Hoffeld |
| 20 | 95273 | Kirchwald |
| 21 | 97169 | Rübenach |
| 22 | 94345, 94342, 94341, 94344, 94343 | Koblenz, Coblenza, Coblenz, Coblence |
| 23 | 95624 | Scheid |
| 24 | 94482 | Barweiler |
| 25 | 98782 | Ettringen |
| 26 | 98047 | Kerschenbach |
| 27 | 96077 | Ochtendung |
| 28 | 95205 | Leudersdorf |
| 29 | 97844 | Wiesbaum |
| 30 | 95629 | Wolken |
| 31 | 95437 | Schüller |
| 32 | 96655 | Welling |
| 33 | 97673 | Ormont |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 34 | 96808 | Nohn |
| 35 | 96218 | Mayen |
| 36 | 99800 | Nachtsheim |
| 37 | 94789 | Lonnig |
| 38 | 100624 | Weiler |
| 39 | 99555 | Berndorf |
| 40 | 96648 | Auw bei Prüm |
| 41 | 95967 | Reuth |
| 42 | 97102 | Kelberg |
| 43 | 100919 | Oberbettingen |
| 44 | 93931 | Lehmen |
| 45 | 100177 | Kehrig |
| 46 | 100610 | Oberehe-Stroheich |
| 47 | 98900 | Waldesch |
| 48 | 97229 | Brey |
| 49 | 99547 | Boxberg |
| 50 | 98683 | Küttig |
| 51 | 100220 | Retterath |
| 52 | 94372 | Kollig |
| 53 | 96866 | Duppach |
| 54 | 95917 | Urmersbach |
| 55 | 97528 | Mützenich |
| 56 | 97311 | Wascheid |
| 57 | 100975 | Hörschhausen |
| 58 | 96828 | Buchet |
| 59 | 96066 | Müllenborn |
| 60 | 99319 | Gamlen |
| 61 | 96560 | Berlingen |
| 62 | 100532 | Laubach |
| 63 | 99414 | Weinsheim |
| 64 | 96696 | Hatzenport |
| 65 | 96564 | Wierschem |
| 66 | 94180 | Nörtershausen |
| 67 | 94471 | Rengen |
| 68 | 99141 | Herscheid |
| 69 | 97062 | Forst (Eifel) |
| 70 | 97121 | Büdesheim |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 71 | 100376 | Winterspelt |
| 72 | 98662 | Ulmen |
| 73 | 97103 | Büscheich |
| 74 | 94338 | Weiler |
| 75 | 98796 | Weinsfeld |
| 76 | 97698 | Greimersburg |
| 77 | 96341 | Ney |
| 78 | 97948 | Kail |
| 79 | 100746 | Mehren |
| 80 | 96970 | Oberstadtfeld |
| 81 | 94597 | Birresborn |
| 82 | 96711 | Demerath |
| 83 | 95557 | Macken |
| 84 | 96037, 100213 | Hersdorf, Nieder-Hersdorf |
| 85 | 95660, 95128 | Gondershausen, Nieder-Gondershausen |
| 86 | 98605 | Gevenich |
| 87 | 99175 | Salm |
| 88 | 100255 | Dörth |
| 89 | 96458 | Biebernheim |
| 90 | 96334 | Lützkampen |
| 91 | 95546 | Seiwerath |
| 92 | 100183 | Valwig |
| 93 | 99828 | Utzenhain |
| 94 | 96086 | Immerath |
| 95 | 97636 | Dohr |
| 96 | 98597 | Heyweiler |
| 97 | 95901 | Bleckhausen |
| 98 | 96575 | Densborn |
| 99 | 98164 | Kinzenburg |
| 100 | 97661 | Eschfeld |
| 101 | 97413 | Pfalzfeld |
| 102 | 94527 | Meisburg |
| 103 | 98171 | Eckfeld |
| 104 | 98093 | Strohn |
| 105 | 100542 | Feuerscheid |
| 106 | 98462 | Mörsdorf |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 107 | 98398 | Damscheid |
| 108 | 93882 | Mesenich |
| 109 | 96323 | Neidenbach |
| 110 | 94449 | Bettenfeld |
| 111 | 96051 | Sankt Aldegund |
| 112 | 96483 | Ebschied |
| 113 | 100160 | Buch |
| 114 | 94070 | Henschhausen |
| 115 | 97704 | Wiebelsheim |
| 116 | 99981 | Dahnen |
| 117 | 96284 | Neurath |
| 118 | 101153 | Grenderich |
| 119 | 97123 | Steinborn |
| 120 | 96003 | Hontheim |
| 121 | 100032 | Oberöfflingen |
| 122 | 94935 | Oberpierscheid |
| 123 | 97319 | Sefferweich |
| 124 | 93976 | Oberweiler |
| 125 | 99765 | Greimerath |
| 126 | 96373 | Blankenrath |
| 127 | 100604 | Oberkail |
| 128 | 97411 | Völkenroth |
| 129 | 93949 | Liebshausen |
| 130 | 95653 | Klosterkumbd |
| 131 | 95927 | Budenheim |
| 132 | 100552 | Fließem |
| 133 | 98113 | Grosslittgen |
| 134 | 100318 | Briedel |
| 135 | 94963 | Karlshausen |
| 136 | 520306 | Oberheimbach |
| 137 | 100390 | Bausendorf |
| 138 | 95646 | Wüschheim |
| 139 | 100137 | Weidingen |
| 140 | 100983 | Neuerburg |
| 141 | 95150 | Dichtelbach |
| 142 | 522623 | Gonsenheim |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 143 | 95701, 95703, 95702, 95700 | Mainz, Magonza, Maguncia, Mayence |
| 144 | 97478 | Schnorbach |
| 145 | 100344 | Wissmannsdorf |
| 146 | 97803 | Pickließem |
| 147 | 96652 | Ellern (Hunsrück) |
| 148 | 101105 | Belg |
| 149 | 100637 | Wackernheim |
| 150 | 96180 | Ürzig |
| 151 | 94962 | Niederkail |
| 152 | 96900 | Wolf |
| 153 | 95346 | Nannhausen |
| 154 | 97389 | Obergeckler |
| 155 | 100636 | Bergweiler |
| 156 | 99401 | Daxweiler |
| 157 | 98099 | Bauler |
| 158 | 95419 | Oberweis |
| 159 | 94398 | Mötsch |
| 160 | 95958 | Seibersbach |
| 161 | 93982 | Altrich |
| 162 | 100621 | Ober Kostenz |
| 163 | 94429 | Raversbeuren |
| 164 | 95747 | Warmsroth |
| 165 | 94765 | Starkenburg |
| 166 | 100818 | Mettendorf |
| 167 | 95885 | Herforst |
| 168 | 95978 | Ockenheim |
| 169 | 93907 | Grosswinternheim |
| 170 | 94226 | Röhl |
| 171 | 97816 | Ravengiersburg |
| 172 | 95744 | Messerich |
| 173 | 95945 | Tiefenbach |
| 174 | 100958 | Rümmelsheim |
| 175 | 94440 | Graach an der Mosel |
| 176 | 98262 | Essenheim |
| 177 | 99928 | Sohren |
| 178 | 94273, 95876 | Noviand, Maring-Noviand |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 179 | 98118 | Bodenheim |
| 180 | 98169 | Körperich |
| 181 | 99466 | Ebersheim |
| 182 | 94251 | Nackenheim |
| 183 | 97772 | Heidweiler |
| 184 | 95708 | Peffingen |
| 185 | 100611 | Pohlbach |
| 186 | 96258 | Nieder-Olm |
| 187 | 97351 | Windesheim |
| 188 | 96276 | Engelstadt |
| 189 | 95598 | Lindenschied |
| 190 | 100129 | Horrweiler |
| 191 | 94820 | Idenheim |
| 192 | 96700 | Laufersweiler |
| 193 | 94915 | Orenhofen |
| 194 | 96733 | Gemünden |
| 195 | 96506 | Münchwald |
| 196 | 96157 | Kleinich |
| 197 | 100795 | Longkamp |
| 198 | 96217 | Biesdorf |
| 199 | 97360 | Gehlweiler |
| 200 | 93952 | Burgen |
| 201 | 95770 | Bretzenheim |
| 202 | 98389 | Rivenich |
| 203 | 100981 | Nieder-Saulheim |
| 204 | 100910 | Kaschenbach |
| 205 | 96059 | Hofweiler |
| 206 | 95986 | Vendersheim |
| 207 | 98927 | Rhaunen |
| 208 | 101133 | Roxheim |
| 209 | 98270 | Wederath |
| 210 | 99071 | Schwabsburg |
| 211 | 94100 | Hahnheim |
| 212 | 100096 | Sprendlingen |
| 213 | 99089 | Föhren |
| 214 | 101043 | Eisenach |
| 215 | 101064 | Stipshausen |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 216 | 99004, 99374 | Neumagen-Dhron, Neumagen |
| 217 | 97143 | Daubach |
| 218 | 99963 | Seesbach |
| 219 | 100713 | Bosenheim |
| 220 | 97014 | Wörrstadt |
| 221 | 97211 | Detzem |
| 222 | 97035 | Hüffelsheim |
| 223 | 98530 | Hellertshausen |
| 224 | 97358 | Hennweiler |
| 225 | 99541 | Waldböckelheim |
| 226 | 97497 | Bischofsdhron |
| 227 | 100461 | Butzweiler |
| 228 | 96429 | Weinolsheim |
| 229 | 99384 | Wickenrodt |
| 230 | 97348 | Olk |
| 231 | 98220 | Horath |
| 232 | 96670 | Weiperath |
| 233 | 98954 | Monzingen |
| 234 | 95383 | Kenn |
| 235 | 94521 | Heidenburg |
| 236 | 96033 | Guntersblum |
| 237 | 95499 | Biebelnheim |
| 238 | 99019 | Mörschied |
| 239 | 99794 | Flonheim |
| 240 | 98026 | Duchroth |
| 241 | 99978 | Wonsheim |
| 242 | 96583 | Altenbamberg |
| 243 | 97240 | Wintersheim |
| 244 | 99330 | Meckenbach |
| 245 | 99034 | Fürfeld |
| 246 | 96329 | Meddersheim |
| 247 | 99204 | Hoxel |
| 248 | 95468 | Sensweiler |
| 249 | 93886 | Mertesdorf |
| 250 | 94509 | Albig |
| 251 | 101007 | Berschweiler bei Kirn |
| 252 | 98039 | Wendelsheim |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 253 | 94729 | Sirzenich |
| 254 | 94748 | Kirschroth |
| 255 | 95840 | Talling |
| 256 | 95592, 95589, 95590, 95591, 95588 | Trier, Treves, Tréveris, Treviri, Trèves |
| 257 | 97643 | Bescheid |
| 258 | 96107 | Mörsfeld |
| 259 | 99151 | Alzey |
| 260 | 94465 | Thomm |
| 261 | 97719 | Vollmersbach |
| 262 | 95969 | Hettenrodt |
| 263 | 100427 | Lauschied |
| 264 | 97986 | Rehborn |
| 265 | 98472 | Hilscheid |
| 266 | 95250 | Niedermoschel |
| 267 | 97394 | Hamm am Rhein |
| 268 | 96353 | Weierbach |
| 269 | 99117 | Bechtheim |
| 270 | 99799 | Kernscheid |
| 271 | 97498 | Hochborn |
| 272 | 96766 | Liersberg |
| 273 | 93955 | Mauchenheim |
| 274 | 95656 | Rascheid |
| 275 | 96660 | Kriegsfeld |
| 276 | 99364 | Oberreidenbach |
| 277 | 94737 | Breitenheim |
| 278 | 98919, 100833 | Mannweiler, Mannweiler-Cölln |
| 279 | 93934 | Kirchenbollenbach |
| 280 | 98179 | Orbis |
| 281 | 95723 | Niedermennig |
| 282 | 101062 | Schmittweiler |
| 283 | 97110 | Börfink |
| 284 | 98581 | Oberhambach |
| 285 | 99008 | Niederbrombach |
| 286 | 100406 | Ober-Flörsheim |
| 287 | 97863 | Bermersheim |
| 288 | 97060 | Fellerich |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 289 | 97217 | Holzerath |
| 290 | 96597 | Reinsfeld |
| 291 | 96848 | Kappeln |
| 292 | 97652 | Damflos |
| 293 | 100810 | Medard |
| 294 | 96431 | Dielkirchen |
| 295 | 96617 | Oberemmel |
| 296 | 96162 | Gauersheim |
| 297 | 98579 | Buhlenberg |
| 298 | 94359 | Wawern |
| 299 | 95194 | Flörsheim-Dalsheim |
| 300 | 99927 | Nittel |
| 301 | 99580 | Lampaden |
| 302 | 95196 | Ruppertsecken |
| 303 | 99446 | Bisterschied |
| 304 | 100948 | Reichenbach |
| 305 | 99725 | Bolanden |
| 306 | 96444 | Nohen |
| 307 | 94658 | Wiesweiler |
| 308 | 100951 | Kell am See |
| 309 | 100483 | Gusenburg |
| 310 | 97690 | Marnheim |
| 311 | 95125 | Worms |
| 312 | 99617 | Nussbach |
| 313 | 97006 | Dannenfels |
| 314 | 99482 | Bubenheim |
| 315 | 96600 | Achtelsbach |
| 316 | 100403 | Hentern |
| 317 | 94933 | Baumholder |
| 318 | 99188 | Heimbach |
| 319 | 94206 | Falkenstein |
| 320 | 95879 | Wincheringen |
| 321 | 100999 | Bierfeld |
| 322 | 96934 | Einöllen |
| 323 | 99803 | Offstein |
| 324 | 99631 | Niederalben |
| 325 | 97578 | Irsch |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 326 | 100225 | Mandern |
| 327 | 101011 | Kahren |
| 328 | 95904 | Göllheim |
| 329 | 97519 | Nohfelden |
| 330 | 98716 | Imsbach |
| 331 | 98258 | Bobenheim-Roxheim |
| 332 | 99854 | Niederkirchen |
| 333 | 96045 | Dennweiler-Frohnbach |
| 334 | 94791 | Gehrweiler |
| 335 | 94202 | Grossniedesheim |
| 336 | 100165 | Kostenbach |
| 337 | 95599 | Serrig |
| 338 | 100939 | Bosen |
| 339 | 93859 | Bedesbach |
| 340 | 96338 | Dirmstein |
| 341 | 97732 | Wolfersweiler |
| 342 | 99897 | Greimerath |
| 343 | 96247 | Morscholz |
| 344 | 97099 | Essweiler |
| 345 | 98116 | Höringen |
| 346 | 97837 | Freisen |
| 347 | 100513 | Kreimbach |
| 348 | 95422 | Münchweiler an der Alsenz |
| 349 | 95649 | Mörsch |
| 350 | 96144 | Kirf |
| 351 | 94113 | Waldhölzbach |
| 352 | 99870 | Schallodenbach |
| 353 | 95574 | Reichweiler |
| 354 | 94502 | Freudenburg |
| 355 | 98465 | Tiefenthal |
| 356 | 98074 | Kirchheim an der Weinstrasse |
| 357 | 100643 | Ramsen |
| 358 | 98360 | Primstal |
| 359 | 97958 | Haschbach am Remigiusberg |
| 360 | 100582 | Föckelberg |
| 361 | 94403 | Neuhemsbach |
| 362 | 98799 | Namborn |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 363 | 99866 | Kollweiler |
| 364 | 94182 | Herchweiler |
| 365 | 100176 | Weisenheim am Berg |
| 366 | 96833 | Lambsheim |
| 367 | 100924 | Saarhölzbach |
| 368 | 98823 | Konken |
| 369 | 99556 | Oberleuken |
| 370 | 100235 | Bardenbach |
| 371 | 98870 | Hausbach |
| 372 | 99905 | Carlsberg |
| 373 | 95337 | Katzweiler |
| 374 | 94595 | Baltersweiler |
| 375 | 100384 | Oggersheim |
| 376 | 95088 | Mehlingen |
| 377 | 99221 | Kallstadt |
| 378 | 98306 | Lindscheid |
| 379 | 100369 | Hoof |
| 380 | 99938 | Wahlen |
| 381 | 100141 | Rehweiler |
| 382 | 99473 | Enkenbach-Alsenborn |
| 383 | 99841 | Tholey |
| 384 | 98443, 97638 | Mitte, Ludwigshafen am Rhein |
| 385 | 97172 | Erlenbach |
| 386 | 98493 | Kottweiler-Schwanden |
| 387 | 98114 | Rodenbach |
| 388 | 95390 | Siegelbach |
| 389 | 99066 | Michelbach |
| 390 | 94297 | Ellerstadt |
| 391 | 95442 | Hardenburg |
| 392 | 95844 | Bethingen |
| 393 | 96435 | Maudach |
| 394 | 97952 | Fischbach |
| 395 | 96473 | Hargarten |
| 396 | 95939 | Steinbach am Glan |
| 397 | 95417 | Ballern |
| 398 | 95386, 99297 | Ramstein-Miesenbach, Ramstein |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 399 | 100327 | Kaiserslautern |
| 400 | 100704 | Wachenheim an der Weinstrasse |
| 401 | 95982 | Niederlinxweiler |
| 402 | 100347 | Altrip |
| 403 | 96596 | Aschbach |
| 404 | 99857 | Urexweiler |
| 405 | 96688, 96394 | Dannstadt-Schauernheim, Dannstadt |
| 406 | 97315, 100482 | Rödersheim-Gronau, Rödersheim |
| 407 | 97900 | Spesbach |
| 408 | 95408 | Hüttersdorf |
| 409 | 95185 | Weidenthal |
| 410 | 94602 | Fürth |
| 411 | 95341 | Kindsbach |
| 412 | 97144 | Haustadt |
| 413 | 96069 | Lebach |
| 414 | 99557 | Schönenberg-Kübelberg |
| 415 | 96598 | Meckenheim |
| 416 | 98850 | Dansenberg |
| 417 | 97225 | Ruppertsberg |
| 418 | 96541 | Waldleiningen |
| 419 | 96030 | Oberesch |
| 420 | 97440 | Wustweiler |
| 421 | 100830 | Münchwies |
| 422 | 98349 | Stennweiler |
| 423 | 97333 | Königsbach |
| 424 | 100273 | Gerlfangen |
| 425 | 95309 | Mölschbach |
| 426 | 97910 | Falscheid |
| 427 | 96064 | Stelzenberg |
| 428 | 94306 | Gimmeldingen |
| 429 | 94514 | Esthal |
| 430 | 96847 | Diefflen |
| 431 | 97068 | Otterstadt |
| 432 | 98212 | Lambrecht (Pfalz) |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 433 | 100034 | Mittelbrunn |
| 434 | 95077 | Wiesbach |
| 435 | 95171 | Iggelheim |
| 436 | 95907 | Martinshöhe |
| 437 | 100008 | Merchweiler |
| 438 | 520319, 100996 | Landsweiler, Landsweiler-Reden |
| 439 | 100860 | Linden |
| 440 | 97835 | Bechhofen |
| 441 | 95167 | Bexbach |
| 442 | 94786 | Niedaltdorf |
| 443 | 99135 | Erbach |
| 444 | 94671 | Niedersalbach |
| 445 | 97012 | Iggelbach |
| 446 | 96588 | Holz |
| 447 | 99802 | Krähenberg |
| 448 | 95806 | Wallerfangen |
| 449 | 99023 | Hermersberg |
| 450 | 94004 | Lachen |
| 451 | 97922, 97921, 97918, 97919, 97920 | Speyer, Espira, Spire, Spiers, Spires |
| 452 | 100667 | Hülzweiler |
| 453 | 94271 | Hanhofen |
| 454 | 98913 | Wallhalben |
| 455 | 94099, 96798 | Spiesen-Elversberg, Spiesen |
| 456 | 96973 | Heltersberg |
| 457 | 97095 | Beeden |
| 458 | 97573 | Maikammer |
| 459 | 98557 | Geinsheim |
| 460 | 95313 | Kirrberg |
| 461 | 97740 | Harthausen |
| 462 | 98110 | Püttlingen |
| 463 | 95236 | Höheinöd |
| 464 | 94828 | Kirkel-Neuhäusel |
| 465 | 94155 | Neuweiler |
| 466 | 97665 | Venningen |
| 467 | 94829 | Battweiler |

Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | Name |
|---|---|---|
| 468 | 100714 | Bous |
| 469 | 96644 | Leimen |
| 470 | 97406 | Rhodt unter Rietburg |
| 471 | 94033 | Altforweiler |
| 472 | 98063 | Freimersheim (Pfalz) |
| 473 | 95528 | Massweiler |
| 474 | 95713 | Mechtersheim |
| 475 | 97038 | Bierbach |
| 476 | 95870 | Ramberg |
| 477 | 95607 | Altenkessel |
| 478 | 95819 | Weingarten (Pfalz) |
| 479 | 101080 | Lingenfeld |
| 480 | 93922 | Burrweiler |
| 481 | 100041 | Überherrn |
| 482 | 96131 | Niederwürzbach |
| 483 | 520303 | Webenheim |
| 484 | 95255 | Essingen |
| 485 | 95474 | Klarenthal |
| 486 | 96156 | Zeiskam |
| 487 | 96590 | Ludweiler-Warndt |
| 488 | 95903 | Nünschweiler |
| 489 | 99562 | Münchweiler an der Rodalb |
| 490 | 97924 | Albersweiler |
| 491 | 97502 | Ommersheim |
| 492 | 97749 | Sankt Arnual |
| 493 | 100398 | Rimschweiler |
| 494 | 98108 | Mittelbach |
| 495 | 100805 | Wilgartswiesen |
| 496 | 98140 | Pirmasens |
| 497 | 97620 | Gersbach |
| 498 | 97150 | Güdingen |
| 499 | 100968 | Hinterweidenthal |
| 500 | 93865 | Hauenstein |
| 501 | 97737 | Eschringen |
| 502 | 97039 | Ottersheim bei Landau |
| 503 | 97469 | Bellheim |
| 504 | 96027 | Wernersberg |

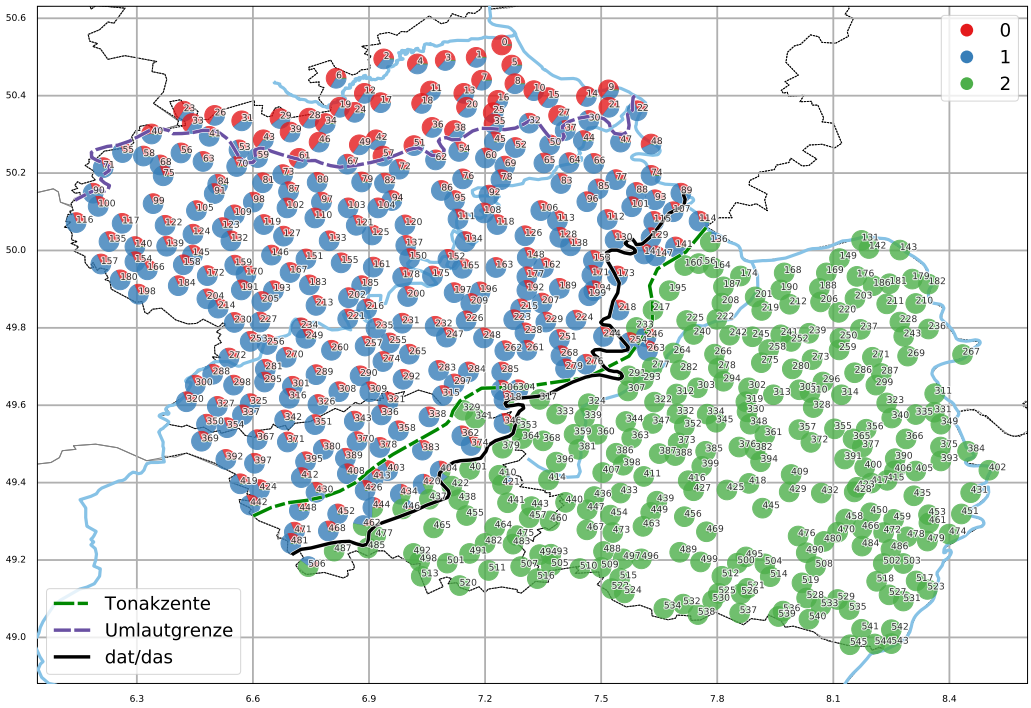
Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 505 | 95537 | Hornbach |
| 506 | 98628 | Lauterbach |
| 507 | 99549 | Böckweiler |
| 508 | 100162 | Ilbesheim bei Landau in der Pfalz |
| 509 | 96198 | Bottenbach |
| 510 | 100455 | Riedelberg |
| 511 | 94727 | Rubenheim |
| 512 | 95118 | Erfweiler |
| 513 | 94530 | Kleinblittersdorf |
| 514 | 98031 | Gossersweiler-Stein |
| 515 | 98451 | Vinningen |
| 516 | 98088 | Brenschelbach |
| 517 | 95583 | Kuhardt |
| 518 | 98806 | Herxheim bei Landau (Pfalz) |
| 519 | 99679 | Klingenmünster |
| 520 | 95197 | Habkirchen |
| 521 | 99684 | Vorderweidenthal |
| 522 | 96538 | Hilst |
| 523 | 97804 | Leimersheim |
| 524 | 99739 | Eppenbrunn |
| 525 | 93857 | Bruchweiler-Bärenbach |
| 526 | 95303 | Erlenbach bei Dahn |
| 527 | 100373 | Hatzenbühl |
| 528 | 97950 | Kapellen-Drusweiler |
| 529 | 100724 | Winden |
| 530 | 95854 | Rumbach |
| 531 | 98366 | Jockgrim |
| 532 | 98771 | Fischbach bei Dahn |
| 533 | 95671 | Dierbach |
| 534 | 100356 | Ludwigswinkel |
| 535 | 98696 | Minfeld |
| 536 | 96765 | Oberotterbach |
| 537 | 96442 | Bobenthal |
| 538 | 101040 | Schönau (Pfalz) |
| 539 | 98210, 94210 | Schweigen-Rechtenbach, Rechtenbach |

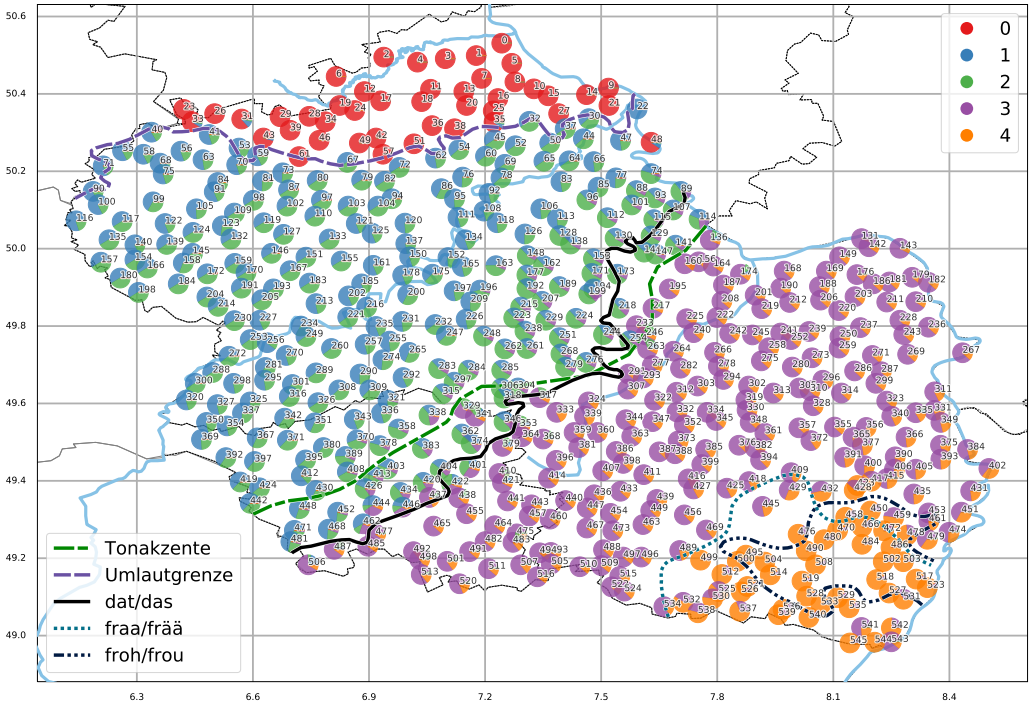
Tabelle A.2: Die Orte im Untersuchungsgebiet des MRhSA.

| ID | GID | NAME |
|---|---|---|
| 540 | 97275 | Steinfeld |
| 541 | 98619 | Büchelberg |
| 542 | 93898 | Hagenbach |
| 543 | 101029 | Neuburg am Rhein |
| 544 | 98156 | Berg (Pfalz) |
| 545 | 100181 | Scheibenhardt |

## A.5 ZUSÄTZLICHE KARTEN ZU KAPITEL 4

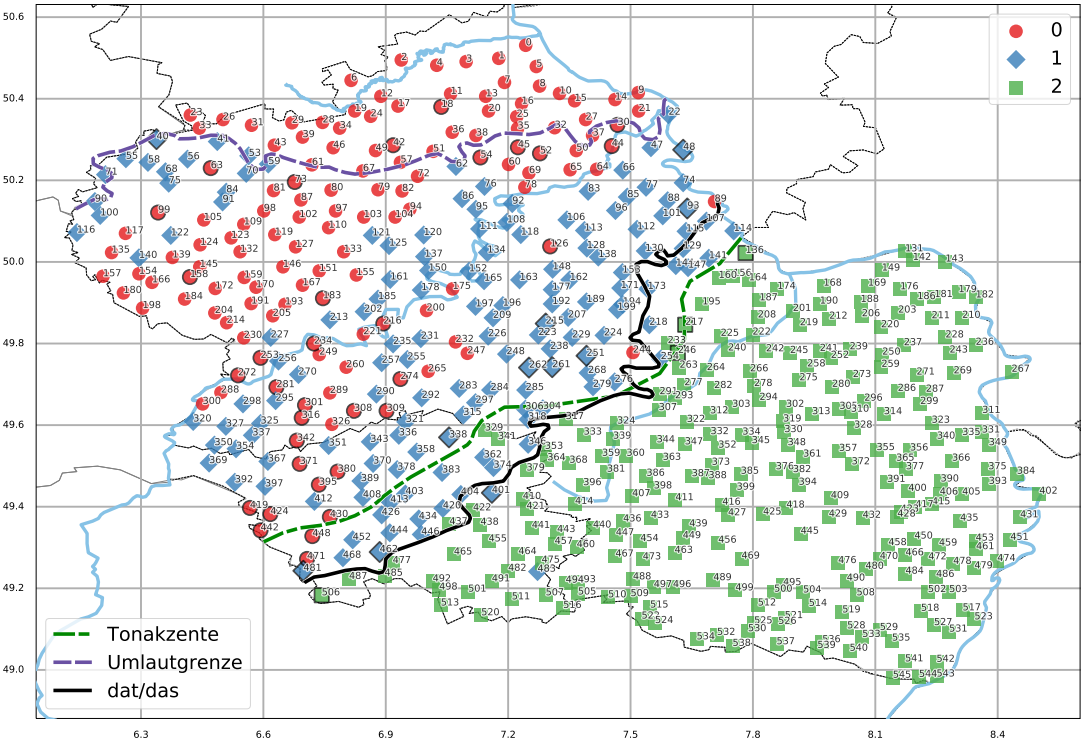



(a) GMM3

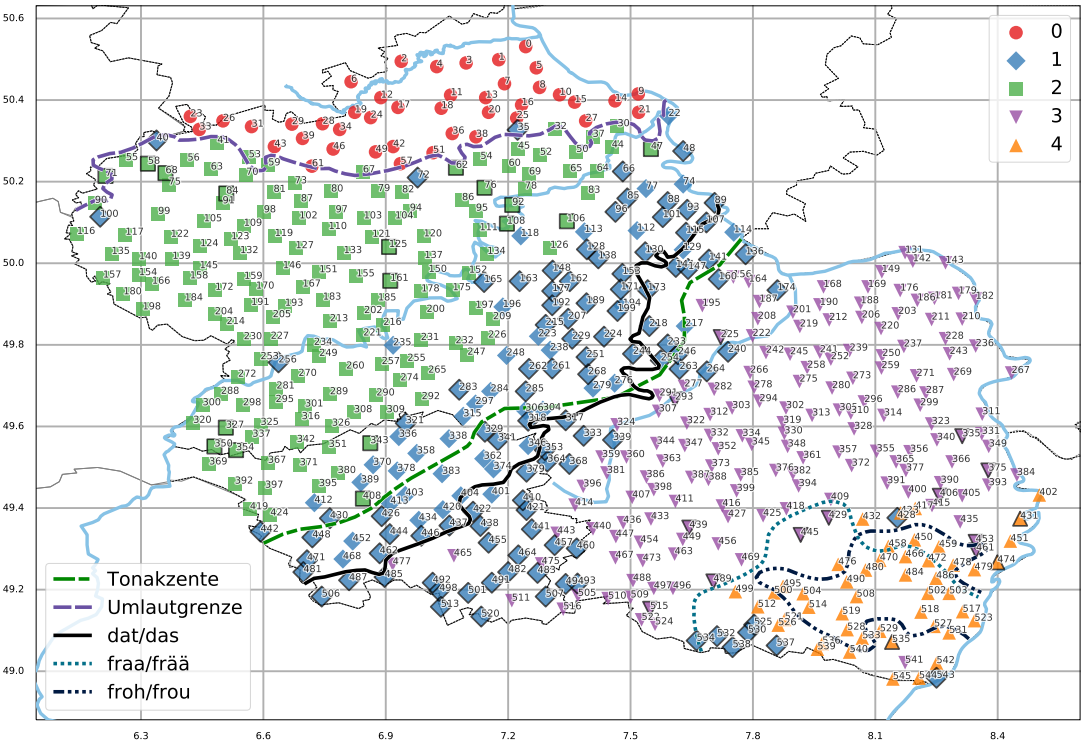



(b) WARD5

Abbildung A.1: Bootstrapping für GMM3 (a) und WARD5 (b) auf dem ALLE-Datenset.

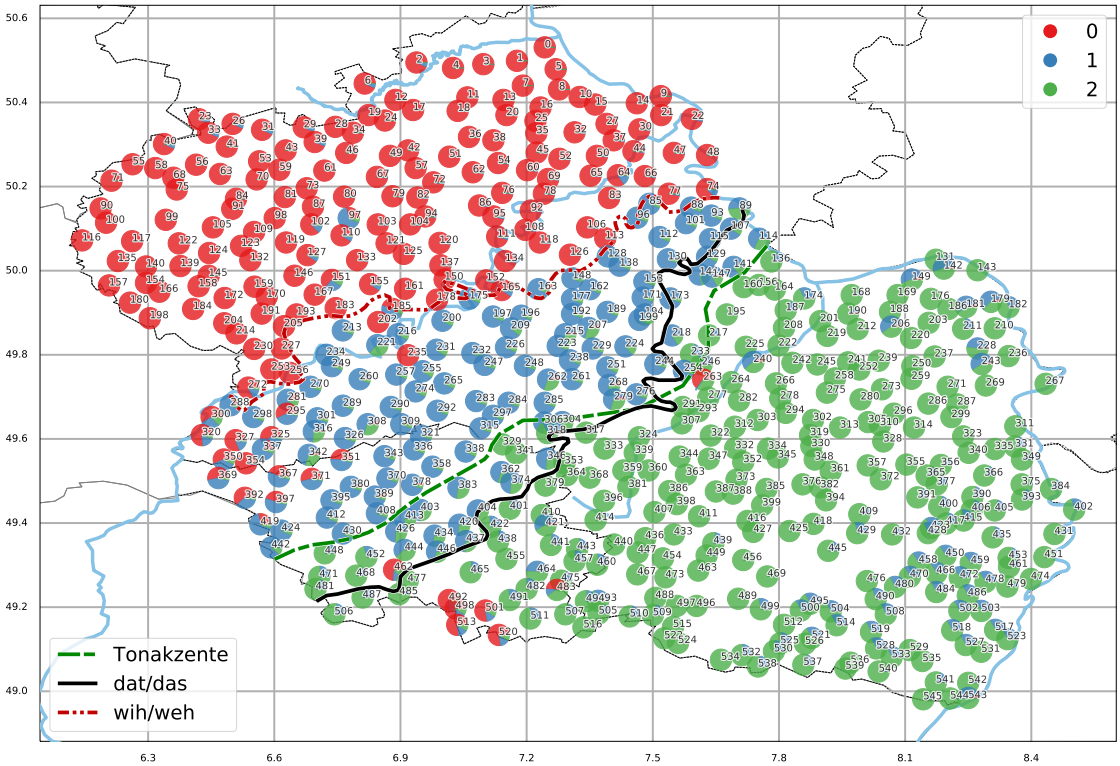



(a) GMM3

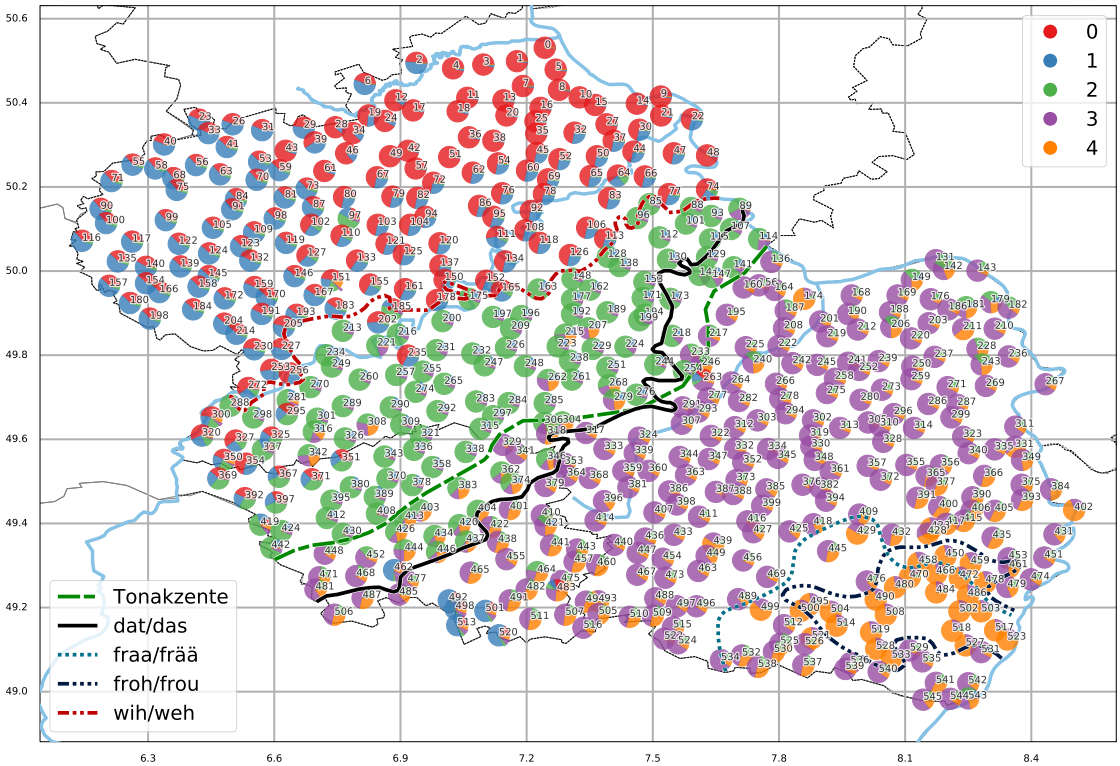



(b) WARD5

Abbildung A.2: Clustering für KMEANS3 (a) und GMM5 (b) auf dem ALLE-Datenset.

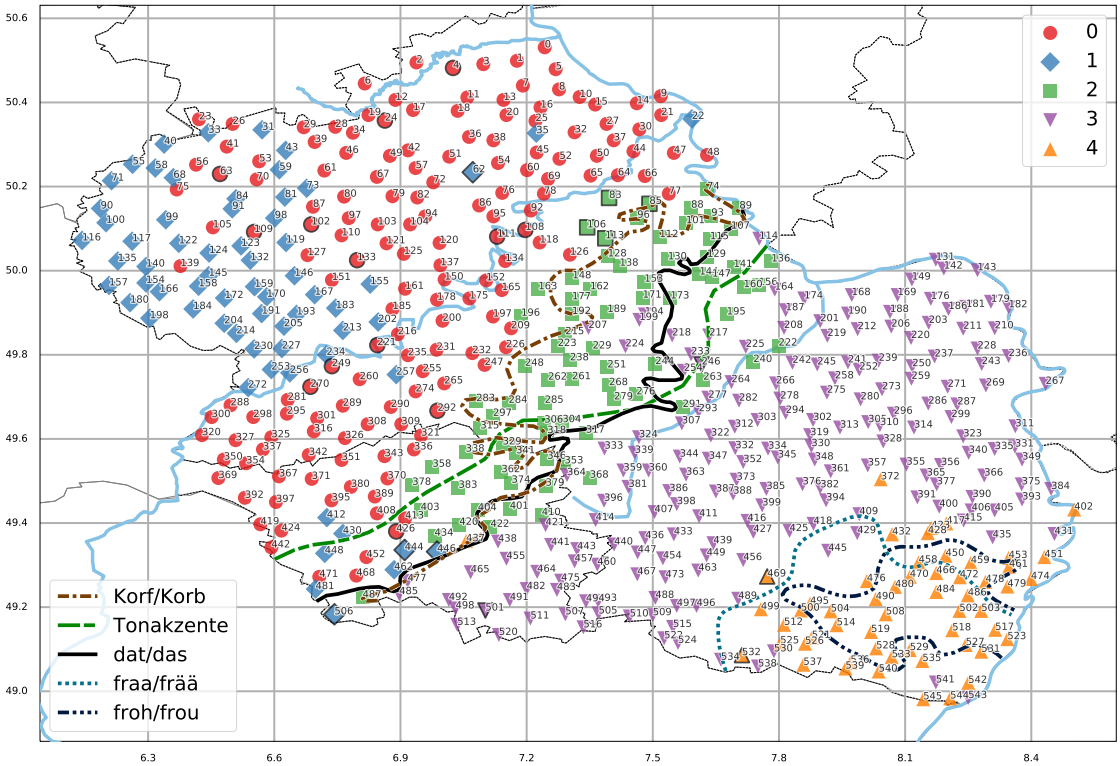



(a) WARD3

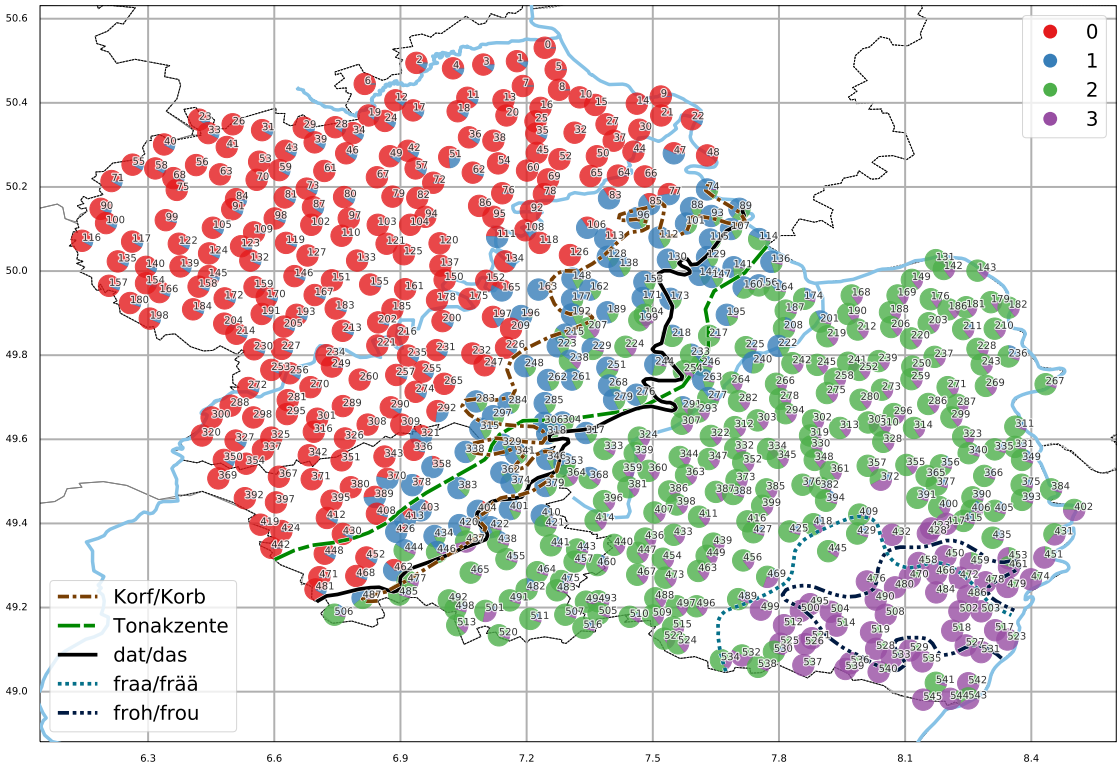



(b) WARD5

Abbildung A.3: Bootstrapping für WARD3 (a) und WARD5 (b) auf dem LANG-Datenset.




(a) KMEANS5-Clustering




(b) Bootstrapping zu KMEANS4

Abbildung A.4: KMEANS5-Clustering (a) und Bootstrapping zu KMEANS4 (b) auf dem WG-Datenset.

## A.6 zusätzliche grafiken zu Kapitel 5

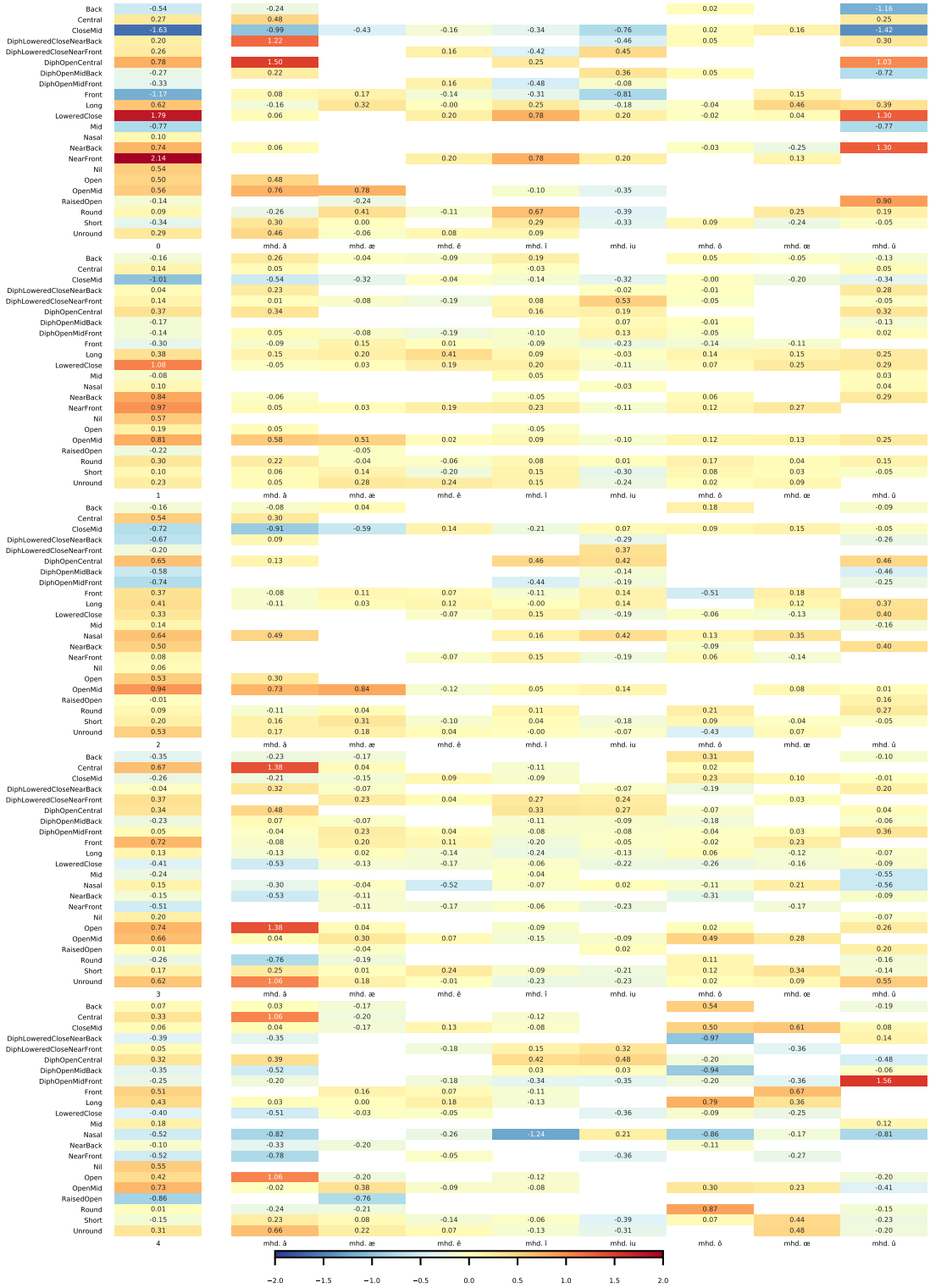



Abbildung A.5: Spektrum der Änderungen in den Clustern nach WARD5 für die Lautklassen der historischen Langvokale.

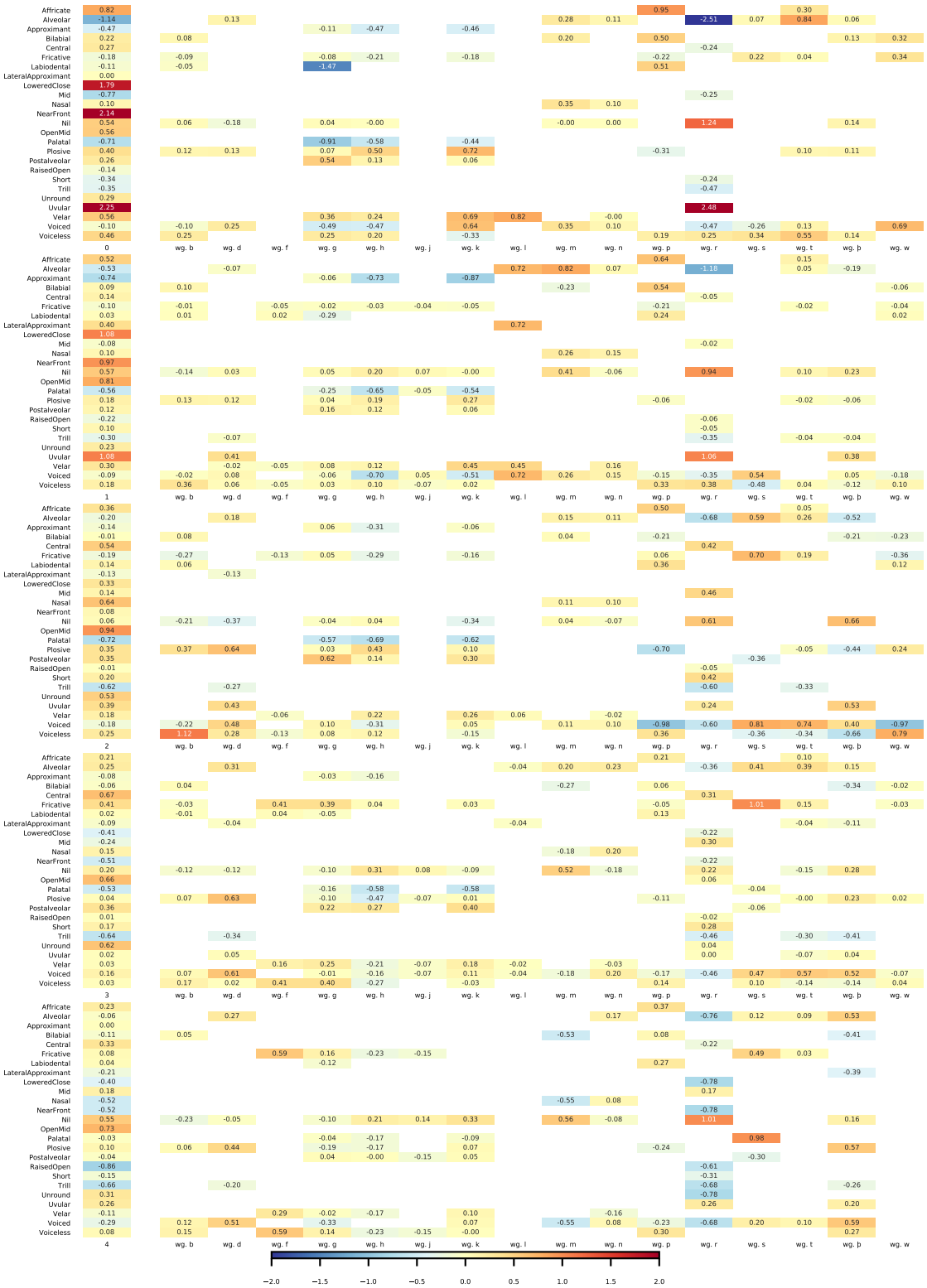



Abbildung A.6: Spektrum der Änderungen in den Clustern nach WARD5 für die Lautklassen der westgermanischen Konsonanten.

# LITERATUR


Abadi, Martín u. a. (2016). „TensorFlow: A System for Large-Scale Machine Learning“. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association, S. 265–283 (siehe S. 64).

Apweiler, Rolf u. a. (2004). „UniProt: the Universal Protein Knowledgebase“. In: *Nucleic Acids Research* 32, S. 115–119 (siehe S. 19).

Aristoteles (1949). *Aristoteloys Analytika : a revised text with introduction and commentary = Aristotles Prior and Posterior Analytics*. Hrsg. von William D. Ross. Clarendon Press (siehe S. 17).

Aurenhammer, Franz und Rolf Klein (2000). „Voronoi Diagrams“. In: *Handbook of computational geometry* 5.10, S. 201–290 (siehe S. 5).

Baader, Franz, Sebastian Brandt und Carsten Lutz (2005). „Pushing the EL Envelope“. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., S. 364–369 (siehe S. 32).

Baader, Franz, Ian Horrocks und Ulrike Sattler (2005). „Description Logics as Ontology Languages for the Semantic Web“. In: *Mechanizing Mathematical Reasoning: Essays in Honor of Jörg H. Siekmann on the Occasion of His 60th Birthday*. Hrsg. von Dieter Hutter und Werner Stephan. Springer, S. 228–248 (siehe S. 30).

— (2008). „Description Logics“. In: *Foundations of Artificial Intelligence* 3, S. 135–179 (siehe S. 21).

Ball, Martin J., Michael Perkins, Nicole Müller und Sara Howard (2008). *The Handbook of clinical Linguistics*. Hrsg. von Martin J. Ball, Michael R. Perkins, Nicole Mller und Sara Howard. Blackwell. 674 S. (siehe S. 38).

Barbiers, Sjef, Hans Bennis, Gunther de Vogelaer, M. Devos und Margreet van der Ham (2005). *Syntactische Atlas van de Nederlandse Dialecten/- Syntactic Atlas of the Dutch Dialects Volume I* (siehe S. 9).

Barbiers, Sjef, Johan Van der Auwera, Hans Bennis, Gunther de Vogelaer und Margreet van der Ham (2008). *Syntactische atlas van de Nederlandse dialecten: volume 2= Syntactic atlas of the Dutch dialects: volume 2* (siehe S. 9).

Beisswanger, Elena, Stefan Schulz, Holger Stenzhorn und Udo Hahn (2008). „BioTop: An Upper Domain Ontology for the Life Sciences“. In: *Applied Ontology* 3.4, S. 205–212 (siehe S. 22).

Bellmann, Günter (1994). *Mittelrheinischer Sprachatlas (MRhSA) : Einführung in den Mittelrheinischen Sprachatlas*. Max Niemeyer Verlag (siehe S. 48, 50, 58).

Bellmann, Günter, Joachim Herrgen und Jürgen Erich Schmidt (1994–2002). *Mittelrheinischer Sprachatlas*. Hrsg. von Georg Drenda. Max Niemeyer Verlag (siehe S. 1, 47).

Berners-Lee, T., R. Fielding und L. Masinter (Aug. 1998). *Uniform Resource Identifiers (URI): Generic Syntax* (siehe S. 26).

Berners-Lee, Tim, James Hendler und Ora Lassila (2001). „The Semantic Web“. In: *Scientific American* 284.5, S. 34–43 (siehe S. 19).

Blancquaert, Edgard und Willem Pée (1925–1982). *Reeks Nederlandse Dialect-Atlassen*. de Sikkel (siehe S. 9).

Bricker, Phillip (2016). „Ontological Commitment". In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University (siehe S. 16).

Brown, Morton B. und Alan B. Forsythe (1974). „Robust Tests for the Equality of Variances". In: *Journal of the American Statistical Association* 69.346, S. 364–367 (siehe S. 68).

Budin, Gerhard, Hans Christian Breuer, Ludwig Maximilian Breuer, Arnold Graf, Barbara Heinisch, Markus Pluschkovits, Rebecca Stocker und Esther Topitz, Hrsg. (2017). *DiÖ (2017): Task-Cluster E: Forschungsplattform*. URL: https://dioe.at/details/ (besucht am 22. 04. 2020) (siehe S. 2).

Butler, Howard, Martin Daly, Allan Doyle, Sean Gillies, S. Hagen und T. Schaub (2016). *The Geojson Format*. Techn. Ber. (siehe S. 52).

Caliński, T. und J. Harabasz (1974). „A Dendrite Method for Cluster Analysis". In: *Communications in Statistics* 3.1, S. 1–27 (siehe S. 76).

Chiarcos, Christian, Sebastian Nordhoff und Sebastian Hellmann (2012). *Linked Data in Linguistics*. Springer (siehe S. 40).

Chomsky, Noam und Morris Halle (1968). „The Sound Pattern of English". In: *International Journal of American Linguistics* 40.1, S. 50–88 (siehe S. 39).

Clements, George N. (1985). „The Geometry of Phonological Features". In: *Phonology* 2.1, S. 225–252 (siehe S. 39).

Clocksin, William F. und Christopher S. Mellish (6. Dez. 2012). *Programming in Prolog: Using the ISO Standard*. Springer Science & Business Media (siehe S. 33).

Cocchiarella, Nino B. (2007). „Formal Ontology and Conceptual Realism". In: *Formal Ontology and Conceptual Realism* 30.4, S. 3–24 (siehe S. 21).

Cortes, Corinna und Vladimir Vapnik (1995). „Support-Vector Networks". In: *Machine Learning* 20.3, S. 273–297 (siehe S. 149).

Cristianini, Nello und John Shawe-Taylor (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge university press (siehe S. 149).

Crystal, David (1985). *A Dictionary of Linguistics and Phonetics*. Blackwell (siehe S. 160).

— (1997). *A Dictionary of Linguistics and Phonology* (siehe S. 38, 162).

Davis, Randall, Howard E. Shrobe und Peter Szolovits (1993). „What Is a Knowledge Representation?" In: *AI Magazine* 14, S. 17–33 (siehe S. 15).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li und Li Fei-Fei (2009). „ImageNet: A Large-scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. IEEE, S. 248–255 (siehe S. 66).

Duerst, M. und M. Suignard (2005). *RFC 3987: Internationalized Resource Identifiers (IRIs)*. RFC 3987 (Proposed Standard), http://www.ietf.org/rfc/rfc3987.txt. Internet Engineering Task Force (siehe S. 26).

Efron, Bradley (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial und Applied Mathematics (siehe S. 79).

Embleton, Sheila (1993). „Multidimensional Scaling as a Dialectometrical Technique: Outline of a Research Project". In: *Contributions to quantitative linguistics*. Springer, S. 267–276 (siehe S. 7).

Falck, Oliver, Alfred Lameli und Jens Ruhose (2018). „Cultural biases in Migration: Estimating non-monetary Migration Costs“. In: *Papers in Regional Science* 97.2, S. 411–438 (siehe S. 13).

Farrar, Scott und D. Terence Langendoen (2003). „A Linguistic Ontology for the Semantic Web“. In: *GLOT international* 7.3, S. 97–100 (siehe S. 37).

— (2009). „An OWL-DL Implementation of Gold“. In: *Linguistic Modeling of Information and Markup Languages*. Springer Netherlands, S. 45–66 (siehe S. 38).

Fellbaum, Christiane (2012). *The Encyclopedia of Applied Linguistics*. Hrsg. von Carol A. Chapelle. John Wiley & Sons, Inc. (siehe S. 37).

Fielding, Roy T. und Richard N. Taylor (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Bd. 7. University of California, Irvine Doctoral dissertation (siehe S. 158).

Fielding, Roy T., Richard N. Taylor, Justin R. Erenkrantz, Michael M. Gorlick, Jim Whitehead, Rohit Khare und Peyman Oreizy (2017). „Reflections on the REST Architectural Style and "Principled Design of the Modern Web Architecture"(Impact Paper Award)“. In: *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ESEC/FSE 2017. Paderborn, Germany: ACM, S. 4–14 (siehe S. 158).

Fielding, Roy, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach und Tim Berners-Lee (1999). *Hypertext Transfer Protocol–HTTP/1.1*. Techn. Ber. (siehe S. 26).

Frings, Theodor (1957). *Grundlegung einer Geschichte der deutschen Sprache*. Bd. 1. Max Niemeyer Verlag (siehe S. 58).

Fujiwara, Yoichi (1976). *A Linguistic Atlas of the Seto Inland Sea: Explanation; a Dialect-geographical Study of the Seto Inland Sea Dialects*. University of Tokyo Press (siehe S. 47).

Gangemi, Aldo, Nicola Guarino, Claudio Masolo, Alessandro Oltramari und Luc Schneider (2002). „Sweetening Ontologies with DOLCE“. In: *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. Springer Berlin Heidelberg, S. 166–181 (siehe S. 22).

Gilliéron, Jules und Edmond Edmont (1902). *Atlas Linguistique de la France*. Champion (siehe S. 2).

Girnth, Heiko (2015). „2. Der Mittelrheinische Sprachatlas (MRhSA) Bidimensionalität und Sprachdynamik“. In: *Regionale Variation des Deutschen*. Hrsg. von Roland Kehrein, Alfred Lameli und Stefan Rabanus. de Gruyter, S. 29–51 (siehe S. 48).

Giunchiglia, Fausto und Ilya Zaihrayeu (2009). „Lightweight Ontologies“. In: *Encyclopedia of Database Systems*. Springer, S. 1613–1619 (siehe S. 20).

Goebl, Hans (1982). *Dialektometrie - Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Verlag der Osterreichischen Akademie der Wissenschaften (siehe S. 2, 3).

— (1984). *Dialektometrische Studien - Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Max Niemeyer Verlag (siehe S. 2, 6).

Goodwin, Robert P. (1965). *Selected Writings of St. Thomas Aquinas*. Bobbs-Merril Co (siehe S. 18).

Gooskens, Charlotte (2004). „Norwegian Dialect Distances Geographically Explained“. In: *Language Variation in Europe. Papers from the Second In-*

*ternational Conference on Language Variation in Europe ICLAVE*. Bd. 2, S. 12–14 (siehe S. 13).

Gracia, Jorge J. E. (1999). *Metaphysics and Its Task: The Search for the Categorcal Foundation of Knowledge*. STATE UNIV OF NY PR. 245 S. (siehe S. 17).

Grau, Bernardo Cuenca, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider und Ulrike Sattler (2008). „OWL 2: The next Step for OWL". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 6.4, S. 309–322 (siehe S. 49).

Graves, Alex und Navdeep Jaitly (2014). „Towards End-to-end Speech Recognition with Recurrent Neural Networks". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, S. II–1764–II–1772 (siehe S. 44).

Gruber, Thomas R. (1995). „Toward Principles for the Design of Ontologies used for Knowledge Sharing?" In: *International Journal of Human-Computer Studies* 43.5-6, S. 907–928 (siehe S. 21).

Guarino, Nicola (1997). „Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction and Integration". In: *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Springer Berlin Heidelberg, S. 139–170 (siehe S. 23).

Guarino, Nicola und Pierdaniele Giaretta (1995). „Ontologies and Knowledge Bases: Towards a Terminological Clarification". In: *Towards very Large Knowledge bases: Knowledge Building and Knowledge sharing*. IOS Press, S. 25–32 (siehe S. 20).

Guarino, Nicola, Daniel Oberle und Steffen Staab (2009). „What Is an Ontology?" In: *Handbook on Ontologies*. Springer Berlin Heidelberg, S. 1–17 (siehe S. 25).

Guha, Ramanathan und Dan Brickley (2014). *RDF Schema 1.1*. W3C Recommendation. http://www.w3.org/TR/2014/REC-rdf-schema-20140225/. W3C (siehe S. 29).

Guyon, I., B. Boser und V. Vapnik (1993). „Automatic Capacity Tuning of Very Large VC-dimension Classifiers". In: *Advances in Neural Information Processing Systems*. Morgan Kaufmann, S. 147–155 (siehe S. 149).

Hall, T. Alan (2011). *Phonologie: Eine Einführung*. de Gruyter (siehe S. 43).

Hartmann, R. R. K. (1973). *Dictionary of Language and Linguistics*. Hrsg. von Francis C. Stork (siehe S. 37, 160).

Hayes, Bruce (2011). *Introductory Phonology*. John Wiley & Sons (siehe S. 40).

Heeringa, Wilbert Jan (2004). „Measuring Dialect Pronunciation Differences using Levenshtein Distance". Diss. (siehe S. 3, 8, 9).

Herre, Heinrich (2010). „General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling". In: *Theory and Applications of Ontology: Computer Applications*. Springer Netherlands, S. 297–345 (siehe S. 22).

— (2013). „Formal Ontology and the Foundation of Knowledge Organization". In: *KNOWLEDGE ORGANIZATION* 40.5, S. 332–339 (siehe S. 25).

Herre, Heinrich, B. Heller, P. Burek, R. Hoehndorf, F. Loebe und H. Michalek (2007). *General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Part I: Basic Principles*. Techn. Ber. Research

Group Ontologies in Medicine (Onto-Med), University of Leipzig (siehe S. 22).

Herrgen, Joachim (2010). „The Linguistic Atlas of the Middle Rhine (MRhSA): A study on the Emergence and Spread of Regional Dialects". In: *Language and Space–An International Handbook of Linguistic Variation* 1, S. 668–686 (siehe S. 53).

Herrgen, Joachim und Jürgen Erich Schmidt (1989). „Dialektalitätsareale und Dialektabbau". In: *Dialektgeographie und Dialektologie.* 60, S. 304–346 (siehe S. 157).

Hitzler, Pascal, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider und Sebastian Rudolph (2009). *OWL 2 Web Ontology Language Primer.* W3C Recommendation. World Wide Web Consortium (siehe S. 31).

Horrocks, Ian (2013). „What Are Ontologies Good For?" In: *Evolution of Semantic Systems.* Springer, S. 175–188 (siehe S. 19).

Horrocks, Ian, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof und Mike Dean (2004). *SWRL: A Semantic Web Rule Language Combining OWL and RuleML.* W3C Member Submission (siehe S. 33).

Horrocks, Ian, Peter F. Patel-Schneider, Sean Bechhofer und Dmitry Tsarkov (2005). „OWL rules: A Proposal and Prototype Implementation". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 3.1, S. 23–40 (siehe S. 33).

Huang, Anna (2008). „Similarity Measures for Text Document Clustering". In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.* Bd. 4, S. 9–56 (siehe S. 4).

Hubert, Lawrence und Phipps Arabie (1985). „Comparing Partitions". In: *Journal of Classification* 2.1, S. 193–218 (siehe S. 78).

Hummel, Lutz (1993). *Dialektometrische Analysen zum Kleinen Deutschen Sprachatlas (KDSA).* Niemeyer (siehe S. 3, 6).

Huson, Daniel H. (1998). „SplitsTree: Analyzing and Visualizing Evolutionary Data." In: *Bioinformatics* 14.1, S. 68–73 (siehe S. 9).

Huson, Daniel H. und David Bryant (2006). „Application of Phylogenetic Networks in Evolutionary Studies". In: *Molecular biology and evolution* 23.2, S. 254–267 (siehe S. 9).

Huson, Daniel H., Regula Rupp und Celine Scornavacca (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications.* Cambridge University Press (siehe S. 79).

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge University Press (siehe S. 1, 43).

Inwagen, Peter van und Meghan Sullivan (2017). „Metaphysics". In: *The Stanford Encyclopedia of Philosophy.* Hrsg. von Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University (siehe S. 17).

Jaberg, Karl, Jakob Jud und Paul Scheuermeier (1928). *Sprach-und Sachatlas Italiens und der Südschweiz.* Bd. 1. Ringier (siehe S. 2).

Jaccard, Paul (1912). „The Distribution of the Flora in the Alpine Zone." In: *New phytologist* 11.2, S. 37–50 (siehe S. 4).

Jain, Anil K. (Juni 2010). „Data Clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8, S. 651–666 (siehe S. 69).

Jain, Anil K. und Richard C. Dubes (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc. (siehe S. 72).

Janssens, Jules (2006). *Ibn Sina and his Influence on the Arabic and Latin World (Variorum Collected Studies)*. Routledge (siehe S. 18).

Jolliffe, I. T. (1986). „Principal Component Analysis and Factor Analysis". In: *Principal component analysis*. Springer New York, S. 115–128 (siehe S. 69).

Keil, Carsten (2017). *Der VokalJäger. Eine phonetisch-algorithmische Methode zur Vokaluntersuchung. Exemplarisch angewendet auf historische Tondokumente der Frankfurter Stadtmundart*. Weidmannsche Verlagsbuchhandlung (siehe S. 44).

Kenstowicz, Michael (2. Dez. 1993). *Phonology Generative Grammar*. Blackwell textbooks in linguistics. John Wiley & Sons. 720 S. (siehe S. 162).

Kessler, Brett (1995). „Computational Dialectology in Irish Gaelic". In: *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. Morgan Kaufmann Publishers Inc., S. 60–66 (siehe S. 8).

Keyser, Samuel Jay und Kenneth N. Stevens (1994). „Feature Geometry and the Vocal Tract". In: *Phonology* 11.02, S. 207–236 (siehe S. 162).

Klepsch, Alfred, Horst Haider Munske und Robert Hinderling (2003). *Sprachatlas von Mittelfranken*. Universitätsverlag Winter (siehe S. 1).

Kohler, Klaus J. (1984). „Phonetic Explanation in Phonology: The Feature Fortis/Lenis". In: *Phonetica* 41.3, S. 150–174 (siehe S. 44).

Krefeld, Thomas, Stephan Lücke und Emma Mages, Hrsg. (2016). *Zwischen traditioneller Dialektologie und digitaler Geolinguistik: Der Audioatlas siebenbürgisch-sächsischer Dialekte (ASD)*. de. Bd. 2. Universitätsbibliothek der Ludwig-Maximilians-Universität München (siehe S. 157).

Kruskal, J. B. (1964). „Multidimensional Scaling by Optimizing Goodness of Fit to a nonmetric Hypothesis". In: *Psychometrika* 29.1, S. 1–27 (siehe S. 80).

Kruskal, Joseph B. (1983). „An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules". In: *SIAM review* 25.2, S. 201–237 (siehe S. 7).

Kumar, Anand, Barry Smith und Daniel D. Novotny (2004). „Biomedical Informatics and Granularity". In: *Comparative and Functional Genomics* 5.6-7, S. 501–508 (siehe S. 22).

Kurath, Hans (1973). *Handbook of the Linguistic Geography of New England*. Ams PressInc (siehe S. 47).

König, Werner (1996–2006). *Sprachatlas von Bayerisch-Schwaben*. Universitätsverlag Winter (siehe S. 11).

König, Werner und Robert Hinderling (1997). *Bayerischer Sprachatlas : 1. Einführung*. Universitätsverlag Winter (siehe S. 1).

Labov, William (1994). *Principles of Linguistic Change, Vol. 1: Internal Factors (Language in Society, No. 20)*. Language in society. Blackwell (siehe S. 47, 50).

Ladefoged, Peter (1988). „Hierarchical Features of the International Phonetic Alphabet". In: *Annual Meeting of the Berkeley Linguistics Society*. Bd. 14. Linguistic Society of America, S. 124–141 (siehe S. 39).

Ladefoged, Peter (1997). „Linguistic Phonetic Descriptions“. In: *The handbook of phonetic sciences*, S. 589–618 (siehe S. 38, 160–162).

— (2000). *Vowels and Consonants : An Introduction to the Sounds of Languages*. Blackwell (siehe S. 160).

Lameli, Alfred (2013). *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. de Gruyter (siehe S. 1, 3, 6, 10, 59, 79).

— (2019). „Areale Variation im Deutschen „horizontal“ Die Einteilung der arealen Varietäten des Deutschen“. In: Bd. 4. de Gruyter Mouton (siehe S. 10).

Lanthaler, Markus, Manu Sporny und Gregg Kellogg (2014). *JSON-LD 1.0*. W3C Recommendation. http://www.w3.org/TR/2014/REC-json-ld-20140116/. W3C (siehe S. 27).

Leśniewski, Stanisław (1929). „Grundzüge eines neuen Systems der Grundlagen der Mathematik“. In: *Fundamenta mathematicae* 14, S. 1–81 (siehe S. 19).

Levenshtein, Vladimir I. (1966). „Binary Codes Capable of Correcting Deletions, Insertions, and Reversals“. In: *Soviet physics doklady*. Bd. 10. 8, S. 707–710 (siehe S. 7).

Linné, Carl von und Johann Friedrich Gmelin (1788). *Caroli a Linné. Systema Naturae per Regna tria Naturae : Secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis*. 6. impensis Georg Emanuel Beer (siehe S. 3, 18).

Lloyd, S. (1982). „Least Squares Quantization in PCM“. In: *IEEE Transactions on Information Theory* 28.2, S. 129–137 (siehe S. 71).

Lowry, Richard (2014a). *Concepts and Applications of Inferential Statistics* (siehe S. 68).

— (2014b). *Concepts and Applications of Inferential Statistics* (siehe S. 90).

Maaten, Laurens van der und Geoffrey Hinton (2008). „Visualizing Data using t-SNE“. In: *Journal of Machine Learning Research* 9, S. 2579–2605 (siehe S. 80).

MacQueen, James u. a. (1967). „Some Methods for Classification and Analysis of Multivariate Observations“. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Bd. 1. 14, S. 281–297 (siehe S. 71).

Maddieson, Ian und Peter Ladefoged (17. Jan. 1996). *Sounds of the Worlds Languages*. John Wiley & Sons. 450 S. (siehe S. 160, 161).

Manola, Frank und Eric Miller (2004). *RDF Primer*. W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/. W3C (siehe S. 25).

Martin, Roland (1914). „Untersuchungen zur Rhein-Moselfränkischen Dialektgrenze“. Marburg, Phil. Diss., 1913. Diss. (siehe S. 59).

McCarthy, John (1968). „Programs with Common Sense“. In: *Semantic Information Processing*. MIT Press, S. 403–418 (siehe S. 19).

McDavid Jr, Raven I. und Raymond K. O'Cain (1980). *Linguistic Atlas of the Middle and South Atlantic States*. Bd. 1. University of Chicago Press (siehe S. 8).

Minsky, Marvin (1988). „A Framework for Representing Knowledge“. In: *Readings in Cognitive Science*, S. 156–189 (siehe S. 19).

Mohr, Georg und Marcus Willaschek (1. Okt. 2010). *Immanuel Kant: Kritik der reinen Vernunft*. de Gruyter. 690 S. (siehe S. 18).

Moon, Todd K. (1996). „The Expectation-maximization Algorithm". In: *IEEE Signal Processing Magazine* 13.6, S. 47–60 (siehe S. 73).

Moran, Steven Paul (2012). „Phonetics Information Base and Lexicon". Diss. University of Washington (siehe S. 40, 157).

Moran, Steven und Michael Cysouw (2018). *The Unicode Cookbook for Linguists: Managing Writing Systems using Orthography Profiles*. Language Science Press (siehe S. 44).

Moran, Steven, Daniel McCloy und Richard Wright, Hrsg. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology (siehe S. 40, 157).

Motik, Boris, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue und Carsten Lutz (2012). *OWL 2 Web Ontology Language: Profiles*. W3C Working Draft. W3C (siehe S. 33).

Nasrabadi, Nasser M. (2007). „Pattern Recognition and Machine Learning". In: *Journal of Electronic Imaging* 16.4, S. 049901 (siehe S. 68).

Nerbonne, John (2007). „Geographic Distributions of linguistic Variation reflect dynamics of Differentiation". In: *Roots: linguistics in search of its evidential base* 96, S. 267 (siehe S. 13).

— (2009). „Data-driven Dialectology". In: *Language and Linguistics Compass* 3.1, S. 175–198 (siehe S. 9).

Nerbonne, John und Wilbert Heeringa (1997). „Measuring Dialect Distance Phonetically". In: *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology* (siehe S. 8).

Nerbonne, John, Wilbert Heeringa und Peter Kleiweg (1999). „Edit Distance and Dialect Proximity". In: *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison* 15 (siehe S. 8).

Nerbonne, John und Peter Kleiweg (2003). „Lexical Distance in LAMSAS". In: *Computers and the Humanities* 37.3, S. 339–357 (siehe S. 9).

Nerbonne, John und Christine Siedle (2005). „Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede". In: *Zeitschrift für Dialektologie und Linguistik* 72.2, S. 129–147 (siehe S. 8, 9).

Nerbonne, John, Rivka Colen, Charlotte Gooskens, Therese Leinonen und Peter Kleiweg (2011). *Gabmap – A Web Application for Dialectology*. Techn. Ber. (siehe S. 9).

Nichols, Johanna und Tandy Warnow (2008). „Tutorial on Computational Linguistic Phylogeny". In: *Language and Linguistics Compass* 2.5, S. 760–820 (siehe S. 10).

Ogden, C. K., Ivor a. Richards und Bronislaw Malinowski (6. Nov. 2013). *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*. Martino Fine Books. 388 S. (siehe S. 16).

Ohala, John J., Catherine P. Browman und Louis M. Goldstein (1986). „Towards an Articulatory Phonology". In: *Phonology Yearbook* 3, S. 219–252 (siehe S. 161).

Øhrstrøm, Peter, Jan Andersen und Henrik Schärfe (2005). „What Has Happened to Ontology". In: *Conceptual Structures: Common Semantics for Sharing Knowledge*. Hrsg. von Frithjof Dau, Marie-Laure Mugnier und Gerd Stumme. Springer Berlin Heidelberg, S. 425–438 (siehe S. 17).

Parmenides (1986). *Vom Wesen des Seienden: die Fragmente, Griechisch und Deutsch*. Hrsg. von Uvo Hölscher. Suhrkamp Verlag AG. 132 S. (siehe S. 17).

Parzen, Emanuel (1962). „On Estimation of a Probability Density Function and Mode". In: *The annals of mathematical statistics* 33.3, S. 1065–1076 (siehe S. 11).

Pease, Adam, Ian Niles und John Li (2002). „The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications". In: *In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, S. 2002 (siehe S. 22).

Pedregosa, F. u. a. (2011). „Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, S. 2825–2830 (siehe S. 82).

Peirce, Charles Sanders (1974). *Collected Papers of Charles Sanders Peirce*. Harvard University Press (siehe S. 19).

Pellegrini, Thomas und Sandrine Mouysset (2016). „Inferring Phonemic Classes from CNN Activation Maps using Clustering Techniques". In: *Annual conference Interspeech (INTERSPEECH 2016)*, 1290pp (siehe S. 44).

Plonsky, Luke, Jesse Egbert und Geoffrey T. Laflair (2015). „Bootstrapping in Applied Linguistics: Assessing its Potential Using Shared Data". In: *Applied Linguistics* 36.5, S. 591–610 (siehe S. 79).

Poli, R. (2001). „The basic Problem of the Theory of Levels of Reality". In: *Axiomathes* 12.3, S. 261–283 (siehe S. 15).

Pompino-Marschall, Bernd (16. Okt. 2009). *Einführung in die Phonetik*. de Gruyter Mouton (siehe S. 43).

Prokić, Jelena, Çağri Çöltekin und John Nerbonne (2012). „Detecting Shibboleths". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. EACL 2012. Avignon, France: Association for Computational Linguistics, S. 72–80 (siehe S. 158).

Pröll, Simon (2015). *Raumvariation zwischen Muster und Zufall: Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben*. Franz Steiner Verlag (siehe S. 3, 11).

Pröll, Simon, Simon Pickl und Aaron Spettl (2014). „Latente Strukturen in geolinguistischen Korpora". In: *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder*. 4, S. 247–258 (siehe S. 12).

Prud'hommeaux, Eric und Gavin Carothers (2014). *RDF 1.1 Turtle*. W3C Recommendation. http://www.w3.org/TR/2014/REC-turtle-20140225/. W3C (siehe S. 27).

Purschke, Christoph (2011). „Regionalsprache und Hörerurteil : Grundzüge einer perzeptiven Variationslinguistik". Teilw. zugl.: Marburg, Universität , Diss., 2010. Diss. (siehe S. 59, 60).

Rand, William M. (1971). „Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66.336, S. 846–850 (siehe S. 78).

Reenen, Pieter T. van, A. C. M. Goeman und J. Taeldeman (2003). *Goeman-Taeldeman-Van Reenen-Project (GTRP), 1985-1995, Phonology & Morphology of Dutch & Frisian Dialects in 1.1 million Transcriptions, Version 2.2 (cd rom)* (siehe S. 9).

Reynolds, D. A. und R. C. Rose (1995). „Robust text-independent Speaker Identification using Gaussian Mixture Speaker Models“. In: *IEEE Transactions on Speech and Audio Processing* 3.1, S. 72–83 (siehe S. 73).

Rosenberg, Andrew und Julia Hirschberg (2007). „V-measure: A conditional entropy-based external Cluster Cvaluation Measure“. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (siehe S. 137).

Rousseeuw, Peter J. (1987). „Silhouettes: A graphical Aid to the Interpretation and Validation of Cluster Analysis“. In: *Journal of Computational and Applied Mathematics* 20.Supplement C, S. 53–65 (siehe S. 75).

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König und Volker Schmidt (2010). „Quantification and statistical analysis of structural similarities in dialectological area-class maps“. In: *Dialectologia et Geolinguistica* 18.1, S. 73–100 (siehe S. 11).

Salvadores, Manuel, Paul R. Alexander, Mark A. Musen und Natalya F. Noy (2013). „BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF“. In: *Semantic Web* 4.3, S. 277–284 (siehe S. 22).

Schmidt, Jürgen Erich (1986). *Die mittelfränkischen Tonakzente : (rheinische Akzentuierung)*. de. Philipps-Universität Marburg (siehe S. 43, 59, 61).

— (2015). „Historisches Westdeutsch und Hochdeutsch. Der Ein-Schritt-Wandel des Langvokalismus“. In: *Sprachwissenschaft* 40.3 (siehe S. 49, 59, 101, 111).

Schmidt, Jürgen Erich (2017). „Vom traditionellen Dialekt zu den modernen deutschen Regionalsprachen“. In: *Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache*. S. 105–143 (siehe S. 10).

Schmidt, Jürgen Erich und Joachim Herrgen (31. Mai 2011). *Sprachdynamik: eine Einführung in die moderne Regionalsprachenforschung*. Schmidt, Erich Verlag. 464 S. (siehe S. 6, 47, 50).

Schmidt, Jürgen Erich, Joachim Herrgen und Roland Kehrein, Hrsg. (2008a). *Neuerhebung der modernen Regionalsprachen des Deutschen*. URL: https://www.regionalsprache.de (besucht am 30. 09. 2017) (siehe S. 47).

— Hrsg. (2008b). *Regionalsprache.de (REDE)*. URL: https : / / www . regionalsprache.de (besucht am 30. 09. 2017) (siehe S. 1, 47).

Schreiber, Guus und Fabien Gandon (2014). *RDF 1.1 XML Syntax*. W3C Recommendation. http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/. W3C (siehe S. 27).

Schreiber, Guus und Yves Raimond (2014). *RDF 1.1 Primer*. W3C Note. http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/. W3C (siehe S. 26).

Seaborne, Andy und Steven Harris (2013). *SPARQL 1.1 Query Language*. W3C Recommendation. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/. W3C (siehe S. 35).

Séguy, Jean (1973). *La Dialectométrie dans l'Atlas Linguistique de la Gascogne*. Société de linguistique romane (siehe S. 2).

Silverman, Bernard W. (1986). *Density estimation for statistics and data analysis*. Bd. 26. CRC press (siehe S. 11).

Simons, Peter (11. Aug. 2000). *Parts: A Study in Ontology*. OXFORD UNIV PR. 408 S. (siehe S. 24).

Sowa, John F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks / Cole (siehe S. 18).

— (2014). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Morgan Kaufmann (siehe S. 15).

Spiekermann, Helmut H., Doris Tophinke, Petra M. Vogel und Claudia Wich-Reif, Hrsg. (2016). *Dialektatlas Mittleres Westdeutschland (DMW)*. URL: http://www.dmw-projekt.de (besucht am 22. 04. 2020) (siehe S. 2).

Spruit, Marco René, Wilbert Heeringa und John Nerbonne (2009). „Associations Among Linguistic Levels". In: *Lingua* 119.11, S. 1624–1642 (siehe S. 9).

Stearns, Michael Q., Colin Price, Kent A. Spackman und Amy Y. Wang (2001). „SNOMED Clinical Terms: Overview of the Development Process and Project Status." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association, S. 662 (siehe S. 19).

Steinhaus, H. (1956). „Sur la Division des Corp Materiels en Parties". In: *Bull. Acad. Polon. Sci* 1, S. 801–804 (siehe S. 71).

Szmrecsanyi, Benedikt (2012). *Grammatical Variation in British English Dialects: A Study in Corpus-based Dialectometry*. Cambridge University Press (siehe S. 3, 13).

Szmrecsanyi, Benedikt und Nuria Hernández (2007). „Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler". In: *FRED-S")". Available online at http://www. freidok. unifreiburg. de/volltexte/2859/. Freiburg: English Dialects Research Group* (siehe S. 13).

Thun, Harald (2001). *L'Atlas Linguistique Diatopique et Diastratique de l'Uruguay*. Université Stendhal-Grenoble III, Centre de dialectologie (siehe S. 47).

Tipping, Michael E. und Christopher M. Bishop (1999). „Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, S. 611–622 (siehe S. 69).

Trudgill, Peter (1974). „Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography". In: *Language in society* 3.2, S. 215–246 (siehe S. 13).

Turing, Alan M. (1937). „On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* 2.1, S. 230–265 (siehe S. 31).

Uschold, Michael und Michael Gruninger (2004). „Ontologies and Semantics for Seamless Connectivity". In: *ACM SIGMOD Record* 33.4, S. 58–64 (siehe S. 20).

Veith, Werner H. (1984). „Kleiner Deutscher Sprachatlas (KDSA). Dialektologische Konzeption und Kartenfolge des Gesamtwerks". In: *Zeitschrift für Dialektologie und Linguistik*, S. 295–331 (siehe S. 6).

Veith, Werner H., Lutz Hummel und Wolfgang Putschke (1984–1999). *Kleiner Deutscher Sprachatlas*. Niemeyer (siehe S. 6).

Vinh, Nguyen Xuan, Julien Epps und James Bailey (2010). „Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". In: *J. Mach. Learn. Res.* 11, S. 2837–2854 (siehe S. 78).

W3C OWL Working Group (Dez. 2012). *OWL 2 Web Ontology Language Document Overview (Second Edition) - W3C Recommendation 11 December 2012*. World Wide Web Consortium (W3C) (siehe S. 31).

*Linguistic Atlas and Survey of Irish Dialects* (1958-1969). Dublin Institute for advanced studies (siehe S. 8).

Ward, Joe H. (1963). „Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, S. 236–244 (siehe S. 72).

Weir, Bruce S. und C. Clark Cockerham (1984). „Estimating F-statistics for the Analysis of Population Structure". In: *evolution* 38.6, S. 1358–1370 (siehe S. 90).

Wenker, Georg (1877). *Das Rheinische Platt*. Bd. 36. Selbstverlag des Verfass. (siehe S. 1, 49).

Wenker, Georg und Ferdinand Wrede (1888–1923). *Der Sprachatlas des Deutschen Reichs*. Handgezeichnet (siehe S. 1, 49).

Wieling, Martijn, Wilbert Heeringa und John Nerbonne (2007). „An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-Van Reenen-Project Data". In: *Taal en Tongval* 59.1, S. 84–116 (siehe S. 9).

Wiesinger, Peter (1983). „Die Einteilung der deutschen Dialekte". In: *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Hrsg. von Werner Besch, Ulrich Knoop, Wolfgang Putschke und Herbert Ernst Wiegand. de Gruyter, S. 807–900 (siehe S. 49, 59, 156).

Wolff, C. (1963). *Preliminary Discourse on Philosophy in General*. Bd. 167. Library of liberal arts. Bobbs-Merrill (siehe S. 18).